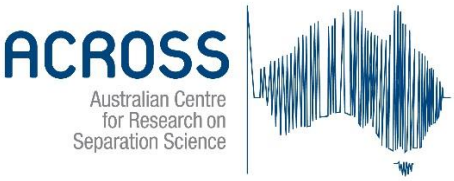
UNIVERSITY
OF TASMANIA

Quantitative Structure-Retention Relationships for Rapid Method Development in Reversed-Phase Liquid Chromatography

Yabin Wen M.Sc.

Submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy



School of Physical Sciences

University of Tasmania

June 2018

Declaration

This thesis contains no material which has been accepted for a degree or diploma by the University or any other institution, except by way of background information and duly acknowledged in the thesis, and to the best of my knowledge and belief no material previously published or written by another person except where due acknowledgement is made in the text of the thesis, nor does the thesis contain any material that infringes copyright.

This thesis may be made available for loan and limited copying and communication in accordance with the Copyright Act 1968.

The publishers of the papers in this thesis hold the copyright for that content, and access to the material should be sought from the respective journals. The remaining non published content of the thesis may be made available for loan and limited copying and communication in accordance with the Copyright Act 1968.

Yabin Wen
May 2018

Statement of co-authorship

The following people and institutions contributed to the publication of work undertaken as part of this thesis:

Candidate: Yabin Wen, ACROSS, School of Physical Sciences, UTAS

Paul R. Haddad, ACROSS, School of Physical Sciences, UTAS

Ruth I.J. Amos, ACROSS, School of Physical Sciences, UTAS

Mohammad Talebi, ACROSS, School of Physical Sciences, UTAS

Robert Shellie, ACROSS, School of Physical Sciences, UTAS

Eva Tyteca, ACROSS, School of Physical Sciences, UTAS

Maryam Taraji, ACROSS, School of Physical Sciences, UTAS

Soo Hyun Park, ACROSS, School of Physical Sciences, UTAS

Roman Szucs, Pfizer Global Research and Development, Sandwich, United Kingdom

John W. Dolan, LC Resources, McMinnville, Oregon, United States

Chris A. Pohl, Thermo Fisher Scientific, Sunnyvale, California, United States

Author details and their roles:

Paper 1, "Towards a chromatographic similarity index to establish localised quantitative structure-retention models for retention prediction: use of retention factor ratio", located in Chapter 3.

E. Tyteca (50%), M. Talebi (7%), R.I.J. Amos (7%), Y. Wen (Candidate, 5%), S.H. Park (5%), M. Taraji (5%), R. Szucs (2%), C.A. Pohl (2%), J.W. Dolan (2%), P.R. Haddad (15%).

- E. Tyteca was the primary author with P.R. Haddad, M. Talebi, and R.I.J. Amos contributed to the idea, its formulation and development.
- P.R. Haddad, M. Talebi and R.I.J. Amos assisted with refinement and presentation.
- Y. Wen (Candidate), S.H. Park, and M. Taraji offered retention data and molecular descriptors for the modelling, along with preliminary results for the pilot study, and offered some descriptions in the experimental section of the manuscript.
- R. Szucs, C.A. Pohl and J.W. Dolan established the need of study and provided feedback on the work.

Paper 2, "Retention time prediction based on molecular structure in pharmaceutical method development: A perspective".

M. Talebi (50%), Robert A. Shellie (7%), R.I.J. Amos (7%), Y. Wen (Candidate, 5%), S.H. Park (5%), M. Taraji (5%), R. Szucs (2%), C.A. Pohl (2%), J.W. Dolan (2%), P.R. Haddad (15%).

- M. Talebi was the primary author with P.R. Haddad, Robert A. Shellie, and R.I.J. Amos contributed to the idea, its formulation and development.
- P.R. Haddad, Robert A. Shellie and R.I.J. Amos assisted with refinement and presentation.
- Y. Wen (Candidate), S.H. Park, and M. Taraji offered retention data and molecular descriptors for the modelling, along with preliminary results for the pilot study, and offered some descriptions in the experimental section of the manuscript.
- R. Szucs, C.A. Pohl and J.W. Dolan established the need of study and provided feedback on the work.

Paper 3, "Retention prediction in reversed phase high performance liquid chromatography using quantitative structure-retention relationships applied to the Hydrophobic Subtraction Model", located in Chapter 4.

Y. Wen (Candidate, 50%), M. Talebi (15%), R.I.J. Amos (15%), R. Szucs (2%), C.A. Pohl (1%), J.W. Dolan (2%), P.R. Haddad (15%).

- Y. Wen was the primary author with P.R. Haddad, M. Talebi, and R.I.J. Amos contributed to the idea, its formulation and development.
- P.R. Haddad, M. Talebi and R.I.J. Amos assisted with refinement and presentation.
- M. Talebi provided general laboratory assistance and GA-PLS algorithm for the modelling.
- R. Szucs, C.A. Pohl and J.W. Dolan established the need of study and provided feedback on the work.

Paper 4, "Retention index prediction using quantitative structure-retention relationships for improving structure identification in Non-Targeted Metabolomics", under revision after favourable review, located in Chapter 5.

Y. Wen (Candidate, 50%), R.I.J. Amos (15%), M. Talebi (15%), R. Szucs (2%), C.A. Pohl (1%), J.W. Dolan (2%), P.R. Haddad (15%).

- Y. Wen was the primary author with P.R. Haddad, M. Talebi, and R.I.J. Amos contributed to the idea, its formulation and development.

- P.R. Haddad, M. Talebi and R.I.J. Amos assisted with refinement and presentation.
- M. Talebi provided general laboratory assistance and GA-PLS algorithm for the modelling.
- R. Szucs, C.A. Pohl and J.W. Dolan established the need of study and provided feedback on the work.

Signed: _____	*Signed:* _____	*Signed:* _____
Paul R. Haddad *Supervisor* *School of Natural Sciences* *University of Tasmania*	*Ruth I.J. Amos* *Co-Supervisor* *School of Natural Sciences* *University of Tasmania*	*Jason Smith* *Head, Discipline of Chemistry* *School of Natural Sciences* *University of Tasmania*

List of publications and presentations

Publications:

1. Y. Wen, R.I.J. Amos, M. Talebi, R. Szucs, C.A. Pohl, J.W. Dolan, P.R. Haddad. "Retention index prediction using quantitative structure-retention relationships for improving structure identification in Non-Targeted Metabolomics", Anal. Chem. 90.15 (2018): 9434-9440. (Chapter 5)

2. Y. Wen, M. Talebi, R.I.J. Amos, R. Szucs, C.A. Pohl, J.W. Dolan, P.R. Haddad. "Retention prediction in reversed phase high performance liquid chromatography using quantitative structure-retention relationships applied to the Hydrophobic Subtraction Model", J. Chromatogr. A 1541 (2018): 1-11. (Chapter 4)

3. E. Tyteca, M. Talebi, R.I.J. Amos, S.H. Park, M. Taraji, Y. Wen, R. Szucs, C.A. Pohl, J.W. Dolan, P.R. Haddad. "Towards a chromatographic similarity index to establish localised quantitative structure-retention models for retention prediction: use of retention factor ratio", J. Chromatogr. A 1486 (2017) 50-58. (Chapter 3)

4. M. Talebi, S.H. Park, M. Taraji, Y. Wen, R.I.J. Amos, P.R. Haddad, R.A. Shellie, R. Szucs, C.A. Pohl, J.W. Dolan. "Retention time prediction based on molecular structure in pharmaceutical method development: a perspective", LCGC North America 34 (8) (2016): 550-558.

Presentations:

5. P.R. Haddad, Y. Wen, R.I.J. Amos, M. Talebi, R. Szucs, C.A. Pohl, J.W. Dolan. "Quantitative structure-retention relationships for retention prediction in RPLC and their application to early stage drug development and non-targeted metabolomics", ISCC 2018, Riva del Garda, Italy. (Oral)

6. Y. Wen, R.I.J. Amos, M. Talebi, R. Szucs, C.A. Pohl, J.W. Dolan, P.R. Haddad. "Retention prediction in reversed-phase high performance liquid chromatography using quantitative structure-retention relationships applied to the Hydrophobic Subtraction Model", RACI R&D Topics 2017, Hobart, Australia. (Oral)

7. Y. Wen, R.I.J. Amos, M. Talebi, R. Szucs, C.A. Pohl, J.W. Dolan, P.R. Haddad. "Retention prediction in reversed phase HPLC using Quantitative Structure-Retention Relationships applied to the Hydrophobic Subtraction Model", HPLC 2017, Prague, Czech Republic. (Oral)

8. Y. Wen, P.R. Haddad, R.I.J. Amos, M. Talebi, R. Szucs, C.A. Pohl, J.W. Dolan, R.A. Shellie. "Similarity Searching in Quantitative Structure-Retention Relationship for Retention

Prediction in Reversed-Phase Liquid Chromatography", ASASS2 2016, Hobart, Australia. (Oral)

9. P.R. Haddad, S.H. Park, M. Taraji, Y. Wen, E. Tyteca, M. Talebi, R.I.J. Amos, R.A. Shellie, R. Szucs, C.A. Pohl, J.W. Dolan. "Prediction of chromatographic retention times based on chemical structures of analytes", ASASS2 2016, Hobart, Australia. (Oral)

10. P.R. Haddad, S.H. Park, M. Taraji, Y. Wen, M. Talebi, R.I.J. Amos, R.A. Shellie, R. Szucs, C.A. Pohl, J.W. Dolan. "Role of Structural Similarity in Prediction of Retention in Reversed-Phase, Ion-Exchange and Hydrophilic Interaction Liquid Chromatography Modes Using Quantitative Structure-Retention Relationships", HPLC 2016, San Francisco, USA. (Oral)

11. P.R. Haddad, S.H. Park, M. Taraji, Y. Wen, E. Tyteca, M. Talebi, R.I.J. Amos, R.A. Shellie, R. Szucs, C.A. Pohl, J.W. Dolan. "Prediction of retention times in reversed-phase, ion-exchange and HILIC modes based on chemical structures", ISC 2016, Cork, Ireland. (Oral)

12. Y. Wen, M. Talebi, R.I.J. Amos, R.A. Shellie, R. Szucs, C.A. Pohl, J.W. Dolan, P.R. Haddad. "Prediction of retention times in reversed-phase liquid chromatography based on chemical structures of analytes", RACI R&D Topics 2015, Melbourne, Australia. (Poster)

Acknowledgements

First of all, I would like to offer this thesis to my unborn child, the biggest surprise in my life. I predicted my 33rd year would be a powerful one, now you are the answer. Hope that one day you will be proud of your father.

Especially, I would like to sincerely thank my primary supervisor, Prof. Paul Haddad, for his guidance and support throughout my study and my life. Being his last PhD student is an honour, and I will never forget his professionalism, humility, and kindness. This thesis would not have been possible without his support and patience.

Additionally, to my beloved wife, Wei Ma, thanks for supporting me and always believing in me. She sacrificed a lot to fulfil my dreams and I will always be grateful. I am also glad to see the achievements you have made for your new career, lovely educator, and I am so proud of you.

I would like to acknowledge my co-supervisors, Dr. Ruth Amos, Dr. Mohammad Talebi, and Assoc/Prof. Robert Shellie, for their excellent guidance, encouragement, and advice on my thesis, publications, and research throughout my PhD. I would also like to acknowledge Dr. Roman Szucs, Dr. John Dolan, and Mr. Chris Pohl, for their helpful discussions upon my publications. Also, many Thanks to Dr. Maryam Taraji and Dr. Soo Hyun Park, for the friendship, discussions, and encouragement.

I also acknowledge the Australian Research Council for the financial support of this research by an ARC Linkage Projects grant (LP120200700) and the Australian Commonwealth Government for providing me the International Postgraduate Research Scholarship (IPRS).

To my dear friends and colleagues, Dr. Min Zhang, Dr. Yan Li, Dr. Feng Li, Shing Chung (John) Lam, Mingxin Liu and Liang Chen. Thanks for sharing the wisdom and experience of life, the good old time will always be in my memory. To the students, post-docs, and staff at ACROSS for offering their assistance and friendship. Also to the members of the old driver football team for organizing the games.

Last but surely not least, I am thankful to my family forever, in particular to my parents, parents-in-law, brother, and brother-in-law, and my cousin, for their silent love, support, and encouragement, I love you all.

Abstract

Reversed-phase liquid chromatography (RPLC) is the most commonly used chromatographic technique in the pharmaceutical industry. In RPLC, a computer-assisted approach is capable of accelerating the process of method development by predicting the retention behaviour of compounds of interest, this would then be followed by an optimisation step to improve chromatographic performance. These objectives can be achieved using a combination of analytical routines and chemometric techniques, and quantitative structure-retention relationship (QSRR) modelling is a promising solution from a variety of chemometric methods. QSRR aims to find meaningful relationships between chromatographic parameters and the molecular descriptors of the compounds of interest. QSRR has been applied for the characterisation of columns, the interpretation of retention mechanisms, the prediction of retention, and the identification of unknown compounds.

The first part of this thesis illustrates the application of QSRR methodology to predict the retention times of compounds extracted from the literature. Several filters including Tanimoto similarity, log D and log P, dual-filtering, and the ratio of retention factors (*k*-ratio) were applied to yield training sets for the construction of QSRR models and the results were explored and compared. QSRR methodology seeks mathematical equations where the chromatographic parameters are expressed as functions of molecular descriptors that were generated, in this case, using Dragon software. To achieve that, the genetic algorithm (GA), employed as feature selection method, combined with a partial least squares (PLS) regression was utilised to correlate measured retention data to various computed molecular descriptors. Among the constructed models, the filter that utilised the ratio of retention factor appeared to be the most effective approach to minimizing prediction errors. But, since the *k*-ratio filter is impractical, it cannot be used directly in QSRR modelling for retention prediction. However, in QSRR modelling the concept of chromatographic similarity should be considered and implemented using a measure of similarity that adequately reflects the retention of compounds. The use of a Tanimoto similarity filter based on the chemical structures of the analytes resulted in unacceptable prediction errors because the level of similarity of compounds in the training sets to the target compounds was not sufficiently high (TS > 0.5). Retention predictions generated using log D, and log P filters, and a dual-filter were not sufficient for the purpose of screening in chromatographic method development. Results indicated the need for larger and more homogenous datasets for QSRR modelling, allowing sufficient numbers of compounds with a high pair-wise similarity to be found when using the Tanimoto filter.

The second part of this thesis demonstrated and compared the contribution of different resources of molecular descriptors to QSRR prediction performance where training sets were formed using different filtering approaches. Molecular descriptors of compounds were calculated using both Dragon, and VolSurf+. Filtering approaches including leave-one-out (LOO), training-test, local compound type (LCT), and local second dominant interaction after hydrophobicity (LSDI) were utilised to allocate compounds to training sets prior to deriving local QSRR models. Instead of predicting retention times directly, retention was predicted indirectly in this chapter by modelling the five solute coefficients of the Hydrophobic Subtraction Model (HSM). Among these four filtering approaches, the LSDI approach showed the best prediction for the five solute coefficients, followed by the LCT approach, demonstrating that approaches embedded with compound classification yielded better prediction of solute coefficients and hence, retention. In terms of the comparison of descriptor resources, no significant difference was observed given the comparable results obtained. The HSM is capable of indirectly providing sufficient accuracy of retention prediction by fitting the predicted solute coefficients and column parameters together.

In the third and main part of the thesis, the utility of the QSRR approach for modelling solute coefficients in the HSM to predict retention times across a wide range of RPLC columns was demonstrated. Different approaches were utilised to cluster compounds for the training sets prior to deriving local QSRR models. The performance of each filtering approach in enhancing prediction accuracy was compared against the classical approach where a global QSRR model is derived using all the compounds in the training set without filtering. The predictive power of the established QSRR models was evaluated using a series of criteria and statistical analysis. In addition, a proof of concept demonstration of the use of QSRR was performed by predicting the retention times of five representative compounds on nine columns for which HS coefficients were known. It was shown that modelling the solute coefficient η' in the HSM is all that is necessary to achieve sufficient prediction accuracy. Results showed improvement in prediction accuracy for the LSDI, LCT and Tanimoto approaches compared with global models derived from the whole dataset without filtering. Of these approaches, the LCT exhibited advantages in its ease of application and the larger number of compounds which can be modelled using this approach. However, the Tanimoto approach was the simplest to apply and, provided that there was a sufficient number of similar compounds in the dataset which meet the TS score cut-off, it yielded sufficiently accurate results. The predictions reported show sufficient accuracy to meet the major objective of this study, namely to determine the likelihood of co-elution of analytes.

In the last part of the thesis, retention index (RI) prediction QSRR models were developed that offer useful predictive ability for compounds having the same molecular weight, allowing false positives to be removed during the interpretation of structure identification in non-targeted metabolomics (NTM). A novel dual-filtering approach combining structural similarity and chromatographic similarity was employed to build suitable training sets for target analytes for the accurate prediction of RI. The elimination of false positives from the list of potential candidate compounds produced by exact mass database searches is demonstrated using the proposed QSRR approach. The predicted RI values generated using the dual-filtering-based QSRR models were very well correlated with the measured data, presenting a percentage root mean square error of prediction (%RMSEP) value of 8.45%. By applying the retention index prediction filter to the modelled compounds in each exact-mass group, 53% of groups were found where at least one false positive could be eliminated. Results have shown that the QSRR modelling using dual-filtering as the strategy for the generation of training sets permits a robust and highly accurate prediction of retention index. Additionally, the results demonstrate that the developed QSRR strategy is capable of eliminating false positives, thereby increasing the confidence of structure identification in MS-based non-targeted metabolomics.

Table of content

List of abbreviations

Acronym	Representation
2D	Two-dimensional
3D	Three-dimensional
ACN	Acetonitrile
ADME	Absorption, distribution, metabolism, and excretion
ANN	Artificial neural networks
B3LYP	Becke 3-parameter (exchange) with correlation by Lee Yang and Parr
CT	Compound type
CV	Cross validation
DAD	Diode array detector
DFT	Density functional theory
F_s	Similarity factor
G126	Global 126
G34	Global 34
GA	Genetic algorithm
GC	Gas chromatography
GETAWAY	Geometry, Topology and Atoms-Weighted AssemblY
HILIC	Hydrophilic-interaction liquid chromatography
HMDB	Human Metabolome Database
HPLC	High-performance liquid chromatography
HRMS	High resolution mass spectrometry
HSM	Hydrophobic-Subtraction Model
IC	Ion chromatography
k	Retention factor
k-ratio	Retention factor ratio
LC	Liquid chromatography
LC-MS	Liquid chromatography-mass spectrometry
LCT	Local compound type
LLD	Local log D
LOO	Leave-one-out

LSDI	Local second dominant interaction
LSS	Linear solvent strength
LTS	Local Tanimoto similarity
LV_s	Latent variables
MAE	Mean absolute error
MD	Molecular descriptor
MIFs	Molecular Interaction Fields
MLR	Multiple linear regression
MOPAC	Molecular Orbital PACkage
MP	Mobile phase
MS	Mass spectrometry
NMR	Nuclear magnetic resonance spectroscopy
NPLC	Normal phase liquid chromatography
NTM	Non-Targeted Metabolomics
OVAT	One-variable-at-a-time
PDA	Photodiode array
PLS	Partial least squares
PM7	Semi-empirical Parametric Method number 7
PQRI	Product Quality Research Institute
QbD	Quality-by-design
QSAR	Quantitative structure-activity relationship
QSPR	Quantitative structure-property relationships
QSRR	Quantitative structure-retention relationship
R^2	Coefficient of determination
RDF	Radial distribution function
REC	Regression Error Characteristic
RI	Retention index
RMSE	Root-mean-square error
RMSEP	Root-mean-square error of prediction
RMSEP%	Percentage root-mean-square error of prediction
RPLC	Reversed-phase liquid chromatography
SDIH	Second dominant interaction after hydrophobicity

SEC	Size exclusion chromatography
SFC	Supercritical fluid chromatography
SMILES	simplified molecular-input line-entry system
SP	Stationary phase
SRD	Sum of ranking difference
t_R	Retention time
t0	Void time
TS	Tanimoto similarity
USP	United States Pharmacopeia Convention
VIP	Variable importance to projection
H	Stationary phase hydrophobicity
S	Stationary phase resistance
A	Stationary phase hydrogen-bond acidity
B	Stationary phase hydrogen-bond basicity
C	Stationary phase ionic interaction
η	Solute hydrophobicity
σ	Solute bulkiness
β	Solute hydrogen-bond basicity
α	Solute hydrogen-bond acidity
κ	Solute ionisation state

1 Introduction

1.1 Thesis overview

Reversed-phase liquid chromatography (RPLC) has been widely used in the separation science community, especially in the field of the pharmaceutical industry. The availability of a broad range of RPLC stationary phases provides opportunities for meaningfully different retention and separation selectivity. With ever more diverse stationary phases available in RPLC, now it is more challenging for chromatographers to choose the most suitable columns for a given set of compounds, or even a starting point for method development. Traditional method development using trial-and-error methods is usually time-consuming and labour-intensive. One option to speed up the process of chromatographic method development is using computer-based methods to predict the retention behaviour of the compounds of interest with good accuracy mainly based on their chemical structure. With the aid of a variety of chemometric approaches, like the quantitative structure-retention relationship (QSRR) methodology, chromatographers have devoted a great deal of effort to propose possible strategies to accelerate method development in RPLC. This thesis aims to explore strategies for the acceleration of method development in RPLC using QSRR methodology to perform computational retention prediction for compounds based solely on their structures. The developed QSRR strategy was successfully used for the retention prediction of new compounds that have never been used in the modelling process, and the elimination of false positives in non-targeted metabolomics (NTM) to improve the confidence of metabolite identification.

This thesis comprises the development of QSRR models for retention prediction and the application of the proposed QSRR in NTM. The retention prediction was performed using three retention databases extracted from the literature, combined with a separate retention database generated from five new compounds and five new columns. The application of the developed QSRR in NTM was conducted on a database of 1882 compounds with known retention indices and molecular weight. To enhance the predictive ability of the QSRR models, the concept of molecular similarity was employed where different filtering approaches were created and utilised as compound classification filters to yield training sets for the construction of QSRR models. The proposed QSRR methodology has been used in column scoping and co-elution prediction in RPLC, as well as the elimination of false positive in NTM.

1.2 Reversed-Phase Liquid Chromatography

As the most widely used separation technology in the pharmaceutical industry, High-Performance Liquid Chromatography (HPLC) has replaced spectroscopic methods and gas

chromatography in numerous quantitative and qualitative analysis applications in the last twenty years [1-3]. In comparison to other methods, a great advantage of HPLC is that it provides a quick, automated and highly accurate method to identify certain chemical components in a sample, combined with over hundreds of stationary phases that are available commercially, enabling the realisation of optimal separation for analytes of interest [4-6]. Additionally, liquid chromatography can be integrated with many types of detector systems like diode detectors, and electrochemical detectors, as well as integration with other systems such as Liquid Chromatography-Mass Spectrometry (LC-MS) and Liquid Chromatography-Nuclear Magnetic Resonance (LC-NMR), another reason for the widespread applicability of this technique [7-10]. RPLC is by far the most popular mode of all types of chromatography which is shown by the fact that nearly 90% of small molecule separations are carried out using RPLC [3, 5, 9, 11].

1.2.1 High-Performance Liquid Chromatography

HPLC, as a technique in the area of analytical chemistry, aims to identify, quantify and separate the components of a mixture [1, 12, 13]. As an integral and major analytical tool in the modern pharmaceutical industry, HPLC has been applied in all stages of method development, drug discovery and high-throughput screening [1-3]. Chromatography, which can be described as a mass transfer process usually involving adsorption, has been given intensive study for decades. HPLC relies on pumps to deliver the mobile phase liquid and the sample mixture through a column filled with adsorbent, leading to the separation of the components of the mixture. The adsorbent, as the active component of the column, is referred to as a "stationary phase", is typically a granular material made of particles of a porous solid (*e.g.* silica, polymers, *etc.*) [1, 14, 15]. Because of the different degrees of interactions with the adsorbent particles, components of the mixture can be separated from each other, leading to the elution of the components as they flow out the column. The pressurised liquid in the HPLC system is typically a mixture of solvents (*e.g.* water, acetonitrile and/or methanol) and is referred to as a "mobile phase" [13, 16]. During the separation process, the temperature and the composition of the mobile phase play a major role by influencing the interactions taking place between the analytes and stationary phase. These interactions are physical in nature, such as hydrophobic (dispersive), dipole–dipole and ionic interactions, and most often, a combination of interactions contributes to the separation [16, 17].

An HPLC instrument typically includes a degasser, sample injector, pumps, and a detector, as can be seen in Figure 1.1 [13, 16, 18, 19]. The degasser, as the solution to the problem of outgassing, is designed to remove the gas from the mobile phase solvents before they are used. The sample injector, as the name suggests, brings the samples into the mobile phase that

carries the components of sample into the column, while the pumps are used to deliver the desired flow and composition of the mobile phase. Various commercially available detectors are in common use now, such as ultraviolet/ visible absorption (UV/Vis), photodiode array (PDA) or detectors based on mass spectrometry. The detector generates a signal proportional to the number of components of the mixture emerging from the stationary phase, hence allowing the quantitative analysis of the analytes [1, 13]. In addition, with the development of the digital microprocessor and embedded software, the control of the HPLC instrument and data processing has been greatly simplified, making the HPLC system more user-friendly even for less experienced chromatographers [18, 20-22].

When separating a given mixture, since there are many combinations of stationary and mobile phases that can be employed, the nature of those phases can be used to classify several types of HPLC techniques [23-25]. Normal-Phase Liquid Chromatography (NPLC), with a solid polar stationary phase and non-polar mobile phase, is usually used to separate polar samples according to the polarity difference of analytes, leading to the least polar components being eluted first, while the most polar components separate last [26, 27]. RPLC, which is the opposite to NPLC, uses a polar liquid mobile phase with a non-polar stationary phase to separate analytes. The most polar components are eluted first, followed by components with lower polarity [1, 6]. Mixtures of water, methanol, and acetonitrile are commonly used as mobile phase, and the non-polar stationary phase can be made by coating silanised silica gel with a non-polar liquid such as a silicone or various hydrocarbons. As an alternative to this type of column, a bonded hydrophobic phase also can be used as the stationary phase, where a hydrophobic molecule is chemically bonded to the polar silica gel. Ion-Exchange or Ion Chromatography (IC), including anion-exchange and cation-exchange, is based on the affinity of the ions and polar molecules to the ion exchanger to separate and determine ions on columns carrying charged functional groups [12, 18]. Other varieties of liquid chromatography such as size exclusion chromatography (SEC) and hydrophilic interaction chromatography (HILIC), *etc*. also have been given intensive study.

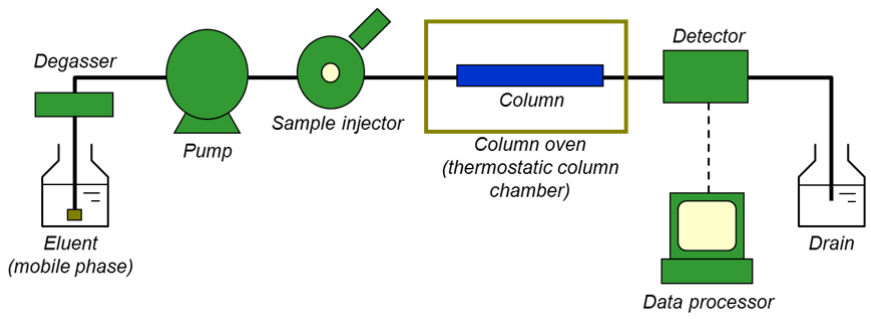


Figure 1.1. Diagram of HPLC system.

1.2.2 Retention mechanism in RPLC

The term "reversed-phase" arises as this mode of chromatography was developed after normal phase chromatography in which a less polar mobile phase is used with a polar stationary phase, thus the mobile phase in RPLC is more polar than the stationary phase [7, 28]. RPLC involves the separation of analytes based on hydrophobicity [29, 30]. In its simplest interpretation, the separation in RPLC mainly depends on the hydrophobic interaction between the hydrophobic nature of the molecule (*e.g*. the carbon backbone) and the non-polar stationary phase ligand (*e.g.* C18) [31, 32]. The hydrophobic interaction is a weak and transient binding force which includes hydrophobic and van der Waals interactions; the more hydrophobic a molecule is, the more retention it has in RPLC [3, 14, 32]. A schematic diagram showing the binding of an analyte to a reversed-phase surface is shown in Figure 1.2. Accordingly, it is reasonable to estimate the elution order based on the solute property of hydrophobicity or hydrophilicity. Figure 1.3 illustrates the general elution order of hydrophilic and hydrophobic analytes.

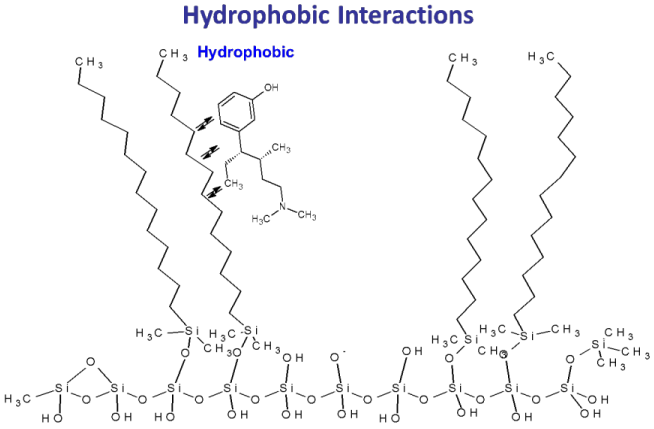


Figure 1.2. Molecular interactions in RPLC.

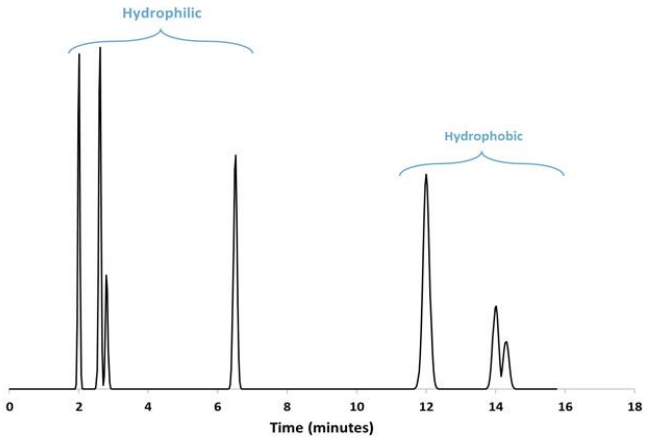


Figure 1.3. Representative reversed-phase chromatogram showing analyte retention order based on hydrophobicity or hydrophilicity.

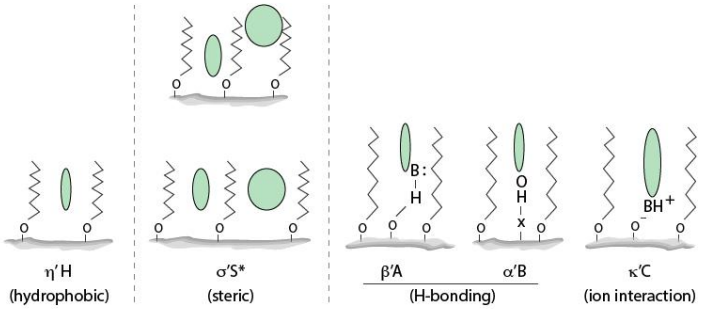


Figure 1.4. Simplified representation of retention processes that correspond to various types of interactions in RPLC. Modified from [32].

It is nowadays generally accepted that the retention of analytes within a RPLC system is caused mainly by hydrophobic interactions between the analyte and the stationary phase [29]. With the intensive study of column characterisation and selectivity in RPLC for decades, other types of interactions between the analytes and the stationary phases which can significantly affect the retention behaviour of solutes have also been found [30, 33-36]. For example, a study performed for a group of alkyl silica phases (*i.e.*, C3-C18) has shown that the retention of compounds on these columns are mainly governed by five different interactions, namely hydrophobicity, steric resistance, hydrogen-bond acidity, hydrogen-bond basicity, and ionic interactions between solute and stationary phase [29, 32, 34-36]. Figure 1.4 provides a simplified representation of these types of interactions.

Given these types of interactions contribute to retention in RPLC, the relative importance of each interaction to solute retention has also been investigated [29, 30, 33]. Wilson and co-workers explored the contribution of each type of interaction in terms of the average change in retention ($\delta \log k$) as a result of a maximum change in the column using a retention database which contained 67 compounds [29, 32]. As expected from the nature of RPLC separation, the hydrophobic interaction contributed most to changes in retention as a result of change in the column. The contribution of the remaining interactions was dependent on the particular compounds and columns [29, 34-36]. Although the remaining types of interaction contribute less than hydrophobicity to overall retention, for column selectivity these terms are still important [29, 37, 38].

1.2.3 Stationary phases in RPLC

For a separation using a liquid chromatography system, the properties of packing materials for the stationary phase are of primary importance [14, 15, 39]. New types of commercial stationary phases for RPLC appear every year with improved stability, efficiency, peak symmetry and selectivity for various analytes [1, 13, 17]. When picking a stationary phase for a separation, the chromatographer should be able to decide which type of column (packed,

capillary, or monolithic column) and what desired characteristics of the column are needed (base material, bonded phase, and bonding density), as even the same types of columns can differ widely in the power of separation across a wide range of manufacturers [6, 7, 19]. In an HPLC system, the most commonly used packing materials are porous particles with average diameters between 3 and 10 μm, and it is highly recommended that 3 μm particle sizes are applied for most pharmaceutical applications. In the middle of the 1900s, small nonporous spherical particles were introduced to increase the efficiency by eliminating dual column porosity [16, 24, 40]. Another attempt to improve the column permeability was achieved by applying a monolithic column. The monolithic column is able to facilitate the accessibility of the adsorbent surface inside the mesopores of the skeleton as it is only 1 μm thick compared to the 5 or even 3 μm particles in conventional packed columns [16, 21].

In RPLC, separations are mainly performed on chemically-modified adsorbents, and analyte interactions with the stationary phase are the primary factors in the process of separation [41, 42]. Silica (SiO_2), the most common substance on the earth, is the most frequently used base material [1, 43]. For the past two centuries, the properties, synthesis, and surface structure of silica have been given intensive study [1, 44]. As the most widely used base material for adsorbents, silica has an array of advantageous properties. Firstly, silica particles are hard enough to withstand harsh packing conditions and the flow of viscous liquids [1, 43]. Additionally, when exposed to different solvents, silica will not shrink or swell. Most importantly, porous silica, as the base material, can provide high surface area which is necessary for a complete separation of analytes. Most of the silica-based HPLC packings are uniform spherical porous particles with narrow particle and pore size distribution[43, 45]. Silica is, however, soluble in a high pH environment. The stability range of silica as a base material has recently been extended to over pH 10 with a chemical modification of a high bonding density of attached alkyl silanes. As an alternative, zirconia is quite stable across a wide pH range (from 1 to 14) [1, 45, 46] and has been suggested as another porous base material in the last decade [1]. But the limitation of zirconia is that it is more difficult to bond functional groups to the surface, which greatly limits its practical applicability. Polymer-based materials also can be used as the base materials and show high pH stability and chemical inertness [13, 16, 47]. Due to the presence of micro porosity in the polymer materials, the application for small molecules separation is somewhat limited. But for size-exclusion chromatography, polymers are the main packing materials [1, 42, 48]. As the most popular RPLC stationary phase, octadecyl carbon chain (C18)-bonded silica columns have been commercially available for decades [41, 49]. In addition, C8-bonded silica, cyano-bonded silica, and phenyl-bonded silica stationary phases are also used widely for the separation in RPLC [1, 17, 50].

Chemical modification is used in the preparation of reversed-phase base materials for the purpose of converting the polar surface into a hydrophobic surface that results in dispersive interactions with the analytes [1, 23, 43]. The conversion of polar silica into a hydrophobic surface needs dense bonding of a thick organic layer, so that the surface of base silica can be shielded effectively [1, 51]. A wide range of different ligands such as C1, C4, C8, C18, Phenyl, Phenyl-hexyl, Nitrile (cyano), and so forth have been tested and bonded on the silica surface [51]. The silanols on the surface can react with different functional groups to form the bonded phase. Practically, almost all of the chromatographic phases that are commercially available are manufactured using a silanisation modification process [14, 15, 47, 50]. There are several types of bonded phases available commercially including alkyl-type phases (C1–C18, C30), phenyl-type phases, and polar embedded stationary phases [1, 15, 31].

Alkyl-type phases are well known to chromatographers since almost 90% of reversed-phase columns are based on these phases [43, 52]. Also, a large number of publications are devoted to the standardisation, characterisation, classification, and comparison of this kind of bonded phase. Among all the investigations, Snyder and co-workers' research attracted a lot of attention and laid the foundation for the future work of column development and application [29-36]. In the book *Practical HPLC Method Development*, Snyder indicates that the retention of non-polar and non-ionic solutes follows the retention pattern: C1 < C4 < C8 $\approx$ C18 in RPLC [16]. Also, dramatic variation of retention for polar and non-polar solutes has been found on the same C18-type columns from different manufacturers even at the same conditions [16, 24, 42, 44].

Phenyl-type bonded phases also have been studied for a long time, since the structure of a phenyl ring on the surface can introduce π–π interactions with some analytes [1, 13, 53]. Compared to alkyl-type bonded phases, phenyl phases show lower methylene selectivity, which means the separation of members of homologous series will be more selective on alkyl phases than phenyl phases [1, 13, 49].

Polar embedded phases were introduced and developed due to the practical need for the separation of highly polar or ionic solutes using highly aqueous mobile phases, because the alkyl or phenyl bonded phases show some drawbacks under these conditions [1, 54]. On one hand, the hydrophobic surface of bonded phases has limited wettability in high aqueous mobile phases [1, 54]. On the other, it was observed that the "phase collapse" effect happens for alkyl chains in pure aqueous mobile phases which greatly decreases the retention of any solutes on the column after a period of use [1, 55, 56]. Polar embedded stationary phases, such as Symmetry-Shield RP18 (Waters) and Supel cosil ABZ (Supelco), contain non-ionised polar groups that are embedded into the train [1, 55, 56]. Another reason that polar groups have

been introduced to the bonded ligands is that these polar groups will inactivate the interactions between silanols and basic analytes by reacting with the residual silanols [1, 57-59].

1.2.4 Mobile phases in RPLC

Hydro-organic mixtures are commonly used as the mobile phases in RPLC [1, 13, 16]. Organic modifiers, such as methanol, acetonitrile and/or combinations of these two are most frequently used [60]. In reversed-phase separation, it is generally accepted that the retention of analytes is governed by the concentration of the organic modifier in the eluent [1, 60, 61]. Therefore, considerations like the compatibility between solvents, the solubility of the sample in the eluent, light transmission, *etc.* need to be taken into account when selecting an appropriate mobile phase for the separation of analytes of interest [1, 59, 62]. For example, methanol and acetonitrile are miscible so this mixture could be used as the mobile phase, while water and dichloromethane are immiscible at most compositions, therefore this mixture should not be used. Similarly, it is well known that a high level of organic solvents should not be used with a high concentration of phosphate buffer since eventually a precipitation will be produced. Additionally, HPLC grade solvents are highly recommended to minimise the contamination of stationary phases and reduce the background absorbance, because impurities exist if the solvent is not purified. Another important consideration for the choice of mobile phase is the light transmission when using UV detection in RPLC. One of the contributing factors to the wide use of acetonitrile as the solvent in reversed-phase separation is because of its low absorbance cut-off wavelength (< 190 nm) [1, 16, 63]. Solvents like acetone and ethyl acetate cannot be used for the separation of analytes at low wavelengths like 210 nm as their UV cut-offs are 330 nm and 256 nm, respectively [1-3]. Like acetonitrile, solvent methanol, ethanol and isopropanol have relatively low UV cut-offs (< 205 nm), but it is always recommended to work at a suitable wavelength with these solvents, for example 210 nm [1-3].

When dealing with separations of the given mixtures using RPLC, besides the type of solvent, parameters such as the eluent composition and the strength of solvent also need to be considered during the process of chromatographic method development [1-3, 64]. RPLC retention also can be illustrated as the effect of the competitive interactions between the components of analytes and the molecules of eluent with the stationary phases [1, 49, 65]. This means that if the interaction between the molecules of eluent and the adsorbent surface of the stationary phases is strong, the interaction between the *analytes* and the adsorbent surface would be weak, thus the analytes will be eluted early, leading to lower retention [10, 66]. Organic solvents, like acetonitrile and methanol, are considered strong solvents in the development of separation methods, solvent strength can be increased by increasing the

proportion of the organic part of mobile phase, allowing early elution of some species of a mixture [1, 64, 67, 68].

Apart from the concentration of the organic solvent in mobile phase, the solvent strength also depends on the type of organic modifier used for the separation [1, 13, 64]. Therefore, the correlation between the concentrations of different organic solvents which are supposed to give similar retention of the analytes of interest has been given intensive study. For the most common organic solvents in RPLC, the ranking based on the solvent strength at the same volume percentage (v/v%) is: tetrahydrofuran > acetonitrile > methanol [1, 13, 64]. According to the different solvent strengths, some general rules have been applied in the process of method development. For example, in order to achieve a similar separation (or elution) on the same stationary phase for a given mixture, compared to an acetonitrile/water eluent, an increased proportion of methanol/water is needed in the mobile phase (a higher concentration of methanol is needed, about 10v/v% more of methanol for every 1v/v% of acetonitrile). Those rules only serve as an approximation as there are more parameters involved in the interactions between the analytes, the solvents, and/or the stationary phases.

Water is usually the base solvent in a reversed-phase application, while other polar solvents such as acetonitrile, methanol, *etc.*, are added in varying proportions [1, 13, 16]. An ideal eluent composition should not affect the selectivity between two species if their ionisation state is independent of the increased organic composition [64, 68, 69]. As we know, for neutral solutes, the selectivity will not be affected by increasing the organic composition. But the ionisation state of ionisable components might be affected when different proportions of organic solvent are used, leading to variations in selectivity [70]. Therefore, buffers are usually utilised to adjust the pH of the aqueous solvent to modify separations, because most pharmaceuticals contain ionisable functionalities such as amino, or carboxylic groups [1, 70]. In fact, the retention of most ionisable compounds is controlled by adjusting the composition and pH of the mobile phase to optimise the separation [16, 24]. It is noticeable that the pH specified for a particular separation is that of the aqueous solvent, and that the addition of organic solvent to the aqueous mixture results in a change in final pH of the mobile phase [1, 13, 16].

1.3 Method development in RPLC

1.3.1 Method development

Three stages need to be considered in the development of a chromatographic method [71, 72]. The first stage is to select the appropriate technique which can provide desired separation selectivity for the analytes of interest [38]. For example, several chromatographic techniques

can be used for the analysis of small-molecule pharmaceuticals, such as liquid chromatography (LC), gas chromatography (GC), supercritical fluid chromatography (SFC) and electrokinetic separations. In addition, the selection of the specific chromatographic technique within one of the above classes is also involved [38, 72]. As we know, within the broad classification of liquid chromatography, RPLC is the most widely used separation technique, but other complementary techniques include hydrophilic-interaction chromatography (HILIC) and ion chromatography (IC) [38, 72-74]. The second stage of method development is column scoping, which aims to identify the most suitable column (stationary phase, SP) offering adequate resolution for the separation of given analytes. This is often found through intensive experimentation [38, 72]. The final stage of method development is to optimise the precise details of the separation conditions, including parameters that are believed to affect the separation, such as the exact mobile phase (MP) composition, column length and temperature, and the flow-rate, *etc* [38, 72]. Basically, according to the properties of the desired analytes and the comparison across the available chromatographic techniques, one would be able to select a starting point to develop a chromatographic method. As the primary phase of method development, the selection of the preferred technique, the stationary phase (or column class) and type of mobile phase (especially the preferred type of organic modifier) is called "*scoping*" (Stages 1 and 2). Further optimisation of details of conditions can be performed *via* a subsequent phase called "*optimisation*" (Stage 3) by implementing some form of experimental design approach [6, 24, 38, 40].

At present, much of the method development in the pharmaceutical industry is carried out by trial-and-error laboratory experimentation [16, 24, 75]. A lot of common parameters need to be considered when HPLC method development is required, such as the selection of the analytical method, sample preparation, the gradient/isocratic elution, column size, the detector, and the wavelength. Traditional approaches like the One-Variable-At-A-Time approach (OVAT) for method development vary one factor at a time while keeping the other factors constant [8, 19, 38]. During this process, all the factors will be examined sequentially until an adequately performing instrumental method is achieved. This process needs to screen different HPLC techniques and a large number of columns (especially for RPLC for which hundreds of columns are available commercially), under numerous conditions, thus it can be a significant waste of resources, as well as requiring excessive consumption of solvents [16, 76]. The process is not only time-consuming and labour-intensive but can lead to chromatographic methods that are not inherently robust and are poorly understood [1, 38].

1.3.2 Column selection

Developing a robust and effective chromatographic method is a diverse and complex process [1, 13]. A comprehensive approach to explore the optimal combination for a separation would be the main investigation of the design space for any proposed analyte mixture and would include the stationary phase, mobile phase, flow-rate, temperature, *etc* [77, 78]. Within the framework of chromatographic method development, choosing an appropriate technique which is expected to provide the desired separation selectivity is the first step, followed by the selection of a suitable stationary phase and the optimisation of details for the separation conditions including mobile phase composition, pH of aqueous solvent, flow-rate, temperature, and so forth [39, 78, 79]. At present, method development in chromatography is still mainly carried out by trial-and-error experimentation, which is not only time-consuming but also inefficient. Moreover, it has become one of the overriding issues in pharmaceutical industry given the fact that a much greater number of drug candidates than at any time in the past have been designed and synthesised [9, 27, 38, 80]. Accordingly, chromatographic method development must evolve to support the needs of the high throughput of drug discovery.

Experienced chromatographers rely on expertise to choose an appropriate type of chromatography for a separation, but they may still struggle with the column selection, which is usually an experimental phase aiming to find the most suitable column (stationary phase) offering adequate resolution for the separation [1, 16, 78]. When selecting a column for the compounds to be analysed, it would be good that chromatographers know what the desired characteristics of a stationary phase are needed, such as base material, bonded phase, and bonding density. Normally, column choices can be narrowed down according to personal experience, literature review, or manufacturer's recommendations, but the basic information provided by the manufacturer regarding the specific column does not allow the choice of a suitable column for a separation directly or even finding an equivalent column with similar separation patterns, given that the same general type of columns could differ greatly in separation power among different manufacturers [3, 5, 81]. This is even further complicated considering the suppliers do not use a standardised testing procedure. Currently, over 700 RPLC stationary phases from various manufacturers are commercially available, and in many cases nominally identical, making it impossible to scope the whole selectivity landscape to choose the optimal stationary phase for a specific separation [62, 72]. Despite intensive study and unprecedented development of packing materials and columns, there is still no consensus on what properties the most suitable stationary phase should have for the selective analysis of diverse sets of analytes [4, 7, 40, 62].

Besides the still repeatedly applied trial-and-error methods, more chromatographic method development is performed by statistical and/or computer-assisted tools [21, 22, 82, 83]. It is well known that time and expenditure could be greatly reduced using an *in silico* approach to investigate chromatographic space [84, 85]. Various types of software packages that assist less experienced chromatographers to understand the mechanisms of separation and to optimise the separation conditions are becoming more attractive [25, 77]. Commercial method development software products such as ACD/LC Simulator (ACD/Labs), DryLab, ChromSword and Osiris, are typically built on initial experimentation with embedded databases to build models either based on molecular-structure-related simulations, retention modelling, or some statistical modelling [38, 72]. While the optimisation phase of method development can be accelerated using software tools, we are still left with the quandary of which column is most suitable for the specific separation of our desired analytes.

Some open resources are very helpful in reducing column choice when replacing a used column for a particular separation. The United States Pharmacopeia Convention (USP) database was designed to assist chromatographers to find HPLC columns equivalent to those that had been used to develop and validate a particular chromatographic procedure [86]. The Product Quality Research Institute (PQRI) database, is capable of selecting an alternative or orthogonal column to a used column by comparing column coefficients [86]. Similarly, a Column Selectivity Database developed by Dwight Stoll and co-workers [87] which contains column parameters for nearly 700 reversed-phase HPLC columns, makes it possible to compare any columns in the database. With the benefit of these open resources, chromatographers can narrow down column selection across a wide range of stationary phases by choosing equivalent and/or orthogonal columns to the column of interest, however choosing the most suitable column for the separation of desired analytes is still not a straightforward process even for experienced chromatographers [4, 78, 81].

1.3.3 Method optimisation

Chromatographic method optimisation is affected by a large number of experimental variables. The optimisation can be achieved in many ways, depending on how many variables are considered [18, 76, 79]. Variables include pressure, temperature, particle size, column length, and eluent velocity. Usually, the OVAT approach is used, which is univariate in nature, to achieve the goal of method optimisation. The OVAT is favoured by non-experts but a drawback of this approach is that it is time-consuming and inefficient, because there is a large number of experiments that must be conducted to determine the effect of each parameter on the retention of the compounds [3, 88]. This OVAT approach requires screening of multiple types of LC techniques and numerous chromatographic columns using a large number of

experimental conditions, which generates significant waste of resources (both human and instrumental) as well as excessive consumption of organic solvents [2, 19, 24, 38]. To maximise the efficiency of scientific discovery and minimise the waste and cost, researchers are keen to perform the fewest experiments that give the most information possible for method optimisation.

Computer-assisted chromatographic optimisation has been intensively used as an alternative [1, 2]. The software packages for optimisation are the most common software tools used for chromatographic method development. These packages run on a premise that the retention of a compound of interest will change in a predictable manner as a function of virtually any continuous chromatographic variable [1, 11]. The most common application of this chromatographic optimisation method is eluent composition, commonly called solvent strength optimisation. For example, with at least two experiments varying the gradient slope for gradient separations or concentration of organic modifier for isocratic separations at a certain temperature, the system then can be used to model any gradient or concentration of organic modifier. Typically, the output from method optimisation package is a resolution map which shows the resolution of the critical pair (two closest eluting peaks) as a function of the parameter(s). The behaviour of certain parameters, including temperature or solvent strength is easily modelled, while other parameters such as the pH or buffer concentration are much more difficult. But, recently there has been renewed interest in automated chromatographic method development [1, 3]. This automated system directly interfaces with the instrument to run or suggest new experiments based on the prior results. Obviously, the advantage of that is the achievement of time-saving in relation to the method development time. However, the lack of maturity of automated optimisation limits its wide application, because this kind of method in principle should follow the logic of chromatographic theory and the mechanism of retention, which unfortunately is not yet fully developed to provide a logical guide [1, 10].

1.4 Quantitative Structure-Retention Relationships

1.4.1 Theory and background

QSRR, as the name suggests, are techniques for relating the variations between compound structures and their retention, and represent a powerful tool in chromatography [10, 89]. QSRR is a technique for relating the variations in one response variable (*Y*-variable) to the variations in several descriptors (*X*-variables), with predictive or explanatory purposes. *Y*-variables are often used as the dependent and *X*-variables as independent variables, Therefore, in QSRR generally *Y*-variables are related to the chromatographic retention of solutes, and *X*-variables encode the molecular structure of solutes [89, 90]. Thus, in chromatography, the principal aim of QSRR is to predict retention data from the molecular structure. QSRR has been applied for

the characterisation of columns by quantitative comparison of separation properties or utilised to provide information for the interpretation of retention mechanisms for various chromatographic conditions (stationary phase, mobile phase, *etc*.) [72, 91, 92]. Additionally, the QSRR method can also offer unique opportunities to predict retention of solutes or to identify analytes [67, 92]. In chromatography, the typical QSRR study comprises several steps: the compilation of a retention database of compounds with known chemical structures, the calculation of molecular descriptors for each structure, a descriptor selection method, QSRR model building, and validation [93, 94]. A scheme of the QSRR methodology is shown in Figure 1.5.

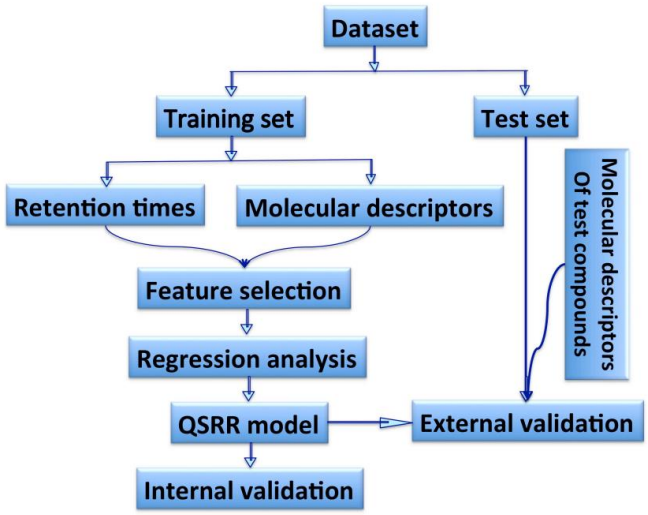


Figure 1.5. Scheme of the QSRR methodology in chromatography.

1.4.2 Molecular descriptors

There are several common ways to represent structures [95], including whole molecule 1D descriptors (sometimes known as 0D), 2D descriptors, and 3D descriptors. 1D descriptors express simple chemical information of a solute such as molecular weight or number of oxygen atoms in the structure, where 2D descriptors are computed from the chemical structure of the solutes of interest when represented by a connection table or a molecular graph. 3D molecular descriptors provide molecular information about the 3D arrangement of structural features and general molecular surfaces and volumes [95-97].

In QSRR modelling, one of the crucial problems is how to represent the molecular structure for QSRR. Usually, molecular descriptors that encode the chemical structures are classified as physico-chemical, quantum-chemical, topological, *etc*. descriptors [95, 96]. One advantage of physico-chemical descriptors is that these descriptors are generally strongly related to the retention of solutes. However, they are often not available or have relatively large errors [89, 90, 93]. Quantum-chemical descriptors provide insights into the mechanism

of chromatographic retention on a molecular level [89, 93] but the correlation to the retention of solutes is often weak, and the calculation is also time-consuming. Topological descriptors are easily generated with present computing tools, but they are not necessarily related to retention phenomena [10, 89].

Computing software like Dragon and VolSurf+ is widely used to generate molecular descriptors based solely on their chemical structures [98-100]. These generated descriptors have been used to evaluate QSRR, quantitative structure-property relationships (QSPR) or quantitative structure-activity relationships (QSAR), as well as for similarity analysis and high-throughput screening of molecule databases [101, 102]. Typically, over 4000 molecular descriptors can be generated using Dragon 6.0 software [98, 103]. The 29 categories of Dragon molecular descriptors are detailed in Table 1.1.

Table 1.1. The categories of molecular descriptors from Dragon and VolSurf+ software

Block ID	Dragon	VolSurf
1	Constitutional descriptors	Size and shape descriptors
2	Ring descriptors	Descriptors of hydrophilic regions
3	Topological indices	Descriptors of hydrophobic regions
4	Walk and path counts	INTEraction enerGY (= INTEGY) moments
5	Connectivity indices	Descriptors of H-bond donor / acceptor regions
6	Information indices	Mixed descriptors
7	2D matrix-based descriptors	Charge State descriptors
8	2D autocorrelations	3D pharmacophoric descriptors
9	Burden eigenvalues	ADME model descriptors
10	P_VSA-like descriptors	
11	ETA indices	
12	Edge adjacency indices	
13	Geometrical descriptors	
14	3D matrix-based descriptors	
15	3D autocorrelations	
16	RDF descriptors	
17	3D-MoRSE descriptors	
18	WHIM descriptors	
19	GETAWAY descriptors	
20	Randic molecular profiles	
21	Functional group counts	
22	Atom-centred fragments	
23	Atom-type E-state indices	
24	CATS 2D	
25	2D Atom Pairs	
26	3D Atom Pairs	
27	Charge descriptors	
28	Molecular properties	
29	Drug-like indices	

Unlike Dragon, where a large number of molecular descriptors are calculated, VolSurf+ software can only generate 128 descriptors for the compounds of interest [99, 100, 104]. VolSurf+ can produce and explore the physico-chemical property space of a molecule (or library of molecules) starting from 3D maps of interaction energies between the molecule and chemical probes (GRID based Molecular Interaction Fields, or MIFs) [99, 105, 106]. One advantage of using VolSurf+ is that it compresses the information present in 3D maps into numerical descriptors optimised for ADME (absorption, distribution, metabolism, and excretion) models and virtual screening, making them are simple to understand and easy to interpret [105, 107]. Those 128 molecular descriptors can be classified into nine categories [100, 108]. The 9 categories of VolSurf+ molecular descriptors are also listed in Table 1.1. In QSRR modelling, chemometric methods are commonly utilised to identify the most suitable subset of molecular descriptors which shows the strongest ability to predict retention times and to build the mathematical relationships [109, 110].

1.4.3 Feature selection and regression analysis

The objective of utilising variable selection methods in QSRR modelling is to use the smallest number of molecular descriptors commensurate with a valid prediction of retention times from among a large number of generated molecular descriptors [38, 67, 72, 92]. A lot of variable selection methods have been elaborated and the proper feature selection is a key to building successful QSRR models. A reason that a proper feature selection method is important in QSRR modelling is because in a given data set some variables may be redundant, irrelevant or represent noise [38, 72]. An good feature selection method is capable of helping to avoid overfitting, reducing the model dimensions, and improving the performance of models [109, 111]. As reported, many feature selection methods like genetic algorithms (GA) and artificial neural networks (ANN) [112, 113] combined with multiple linear regression (MLR) [114] or partial least squares (PLS) [71, 111, 115], have been given intensive attention to build final model in QSRR studies.

As a statistical tool that is commonly used in a QSRR study, MLR has been used widely to handle the selection of molecular descriptors for the construction of QSRR models [10, 89]. With the significant increase in the number of molecular descriptors that can be computed, some new chemometric modelling techniques have been introduced to QSRR modelling in order to manage the greater number of descriptors [109, 116, 117]. PLS is a linear, multiple regression method and it has been used frequently in chemometric and multivariate calibration studies. PLS is particularly useful in handling databases with a large number of variables compared to the number of objects and in the presence of co-linear, redundant, and noisy variables [71, 72, 91]. A PLS method can be expressed as

$$y = a_1 LV_1 + a_2 LV_2 + \cdots + a_m LV_m \qquad 1.1$$

where *y* is the dependent variable, a_1, a_2, ⋯, a_m are the regression coefficients, and LV_i is the *i*-th latent variable. As can be seen from Eq. 1.1, PLS summarises the variation in the independent variables into a small set of linear, orthogonal, and latent variables (LVs) by maximising the covariance between descriptors and the dependent variable [111, 115, 118]. In addition, over-fitting in the models can be minimised by optimising the number of LVs.

1.4.4 Model validation

In QSRR modelling, a training set is used to build QSRR models, and a test set is needed for validation. For this purpose, the measured retention data of the test compounds is extracted and compared with the predicted retention data calculated from the derived QSRR models [67, 72, 73, 102]. The statistical reliability of the formed QSRR models needs to be validated, and this can be performed by several approaches. The coefficient of determination (R^2), the slope of the regression with no forced intercept, the mean absolute error (MAE) and the root-mean-square error of prediction (RMSEP) are commonly used to evaluate the fitness and the predictive ability of the constructed QSRR models [72, 73, 91]. Additionally, the percentage root-mean-square error of prediction (RMSEP%) of retention time for the test set is also a frequently used error reporting method for external validation of the accuracy of QSRR models generated from the training sets.

1.4.5 QSRR accuracy

In many cases, the precision and accuracy of the QSRR models is low, but may still be useful for the interpretation of the retention mechanisms, or the optimisation for the separation of complex mixtures, or the preparation of experimental designs [10, 89, 90, 93]. The predictive accuracy of the QSRR models can be influenced by a number of factors: (i) the feature selection method employed to choose the most informative descriptors, (ii) the modelling approach used to build QSRR models, (iii) the model validation approach utilised, (iv) the number of molecular descriptors incorporated into the QSRR models, (v) the geometry optimisation method used, (vi) the size of the dataset employed in the study, and (vii) the range of diversity or similarity of the molecular structures or characteristics.

In terms of the modelling approach employed for the construction of QSRR models, compound classification may provide greater predictive ability compared with the QSRR models derived from a diverse dataset [119]. As reported [119], compound classification has been achieved based on the log D profile similarity of compounds in a database and the performance of the subset-specific models was improved compared with a QSRR model using no compound classification. Another example can be found from the work by Muteki and co-

workers [120], using a compound-classification-based QSRR methodology to improve the retention time predictability compared with the global models.

1.4.6 Molecular similarity

As the name suggests, structurally similar molecules are more likely to exhibit similar properties [121, 122]. From this, the interest has been increased for the prediction of properties for compounds based on molecular similarity [123, 124]. Compared to a diverse training set, a much more structurally similar subset of compounds in a training set could be generated using this concept and is likely to produce better prediction results. The degree of structural similarity between two compounds can be calculated with the assistance of some chemometric tools, allowing a similarity coefficient to be obtained [125].

The Tanimoto coefficient, as the most commonly used similarity measurement of compounds, appears to be the gold standard in computing the fingerprint-based similarity used in QSRR or QSAR modelling [74, 91, 122, 125]. The Tanimoto coefficient for molecules A and B can be calculated using Eq. 1.2:

$$S_{A,B} = \frac{c}{a + b - c} \qquad 1.2$$

Where, a and b are the bit sets in the fingerprints for A and B, and c is the bit set in common between the two fingerprints. The Tanimoto coefficient takes values between zero and unity, with 0 corresponding to no bits in common and 1 to identical fingerprints [121, 126]. In this thesis, the Tanimoto similarity was employed as a basic filter to select structurally similar compounds to the target compound to form a training set to be used for the subsequent construction of the QSRR models.

1.5 Hydrophobic-Subtraction Model

1.5.1 Theory and background

The Hydrophobic-Subtraction Model (HSM) was originally developed to describe column selectivity in RPLC [29, 30, 33]. Wilson and co-workers used a retention database of 67 diverse compounds on 10 different C_{18} columns to derive a six-term equation for the correlation of retention as a function of solute and column [29, 30, 33]. As can be seen from the equation below (Eq. 1.3), the relative retention of a given compound is defined through a linear combination of five terms which represent the hydrophobicity ($\eta'\mathbf{H}$), steric resistance ($\sigma'\mathbf{S}^*$), hydrogen-bond acidity ($\beta'\mathbf{A}$), hydrogen-bond basicity ($\alpha'\mathbf{B}$), and ionic interactions ($\kappa'\mathbf{C}$), respectively.

$$\log \alpha \equiv \log\left(\frac{k}{k_{\mathrm{EB}}}\right) = \eta'\mathbf{H} - \sigma'\mathbf{S}^* + \beta'\mathbf{A} + \alpha'\mathbf{B} + \kappa'\mathbf{C} \qquad 1.3$$

Where α is the chromatographic selectivity, k is the retention factor of the solute, and k_{EB} is the retention factor of the ethylbenzene. Each term in this equation corresponds respectively to each of the interactions mentioned above. Values of column coefficients for RPLC phases do not vary much under changing mobile phase composition, except that the C term increases with an increase in mobile phase pH. According to Wilson and co-workers, the HSM assumes that the major contribution of hydrophobicity to RPLC retention is subtracted first, in order to better see the remaining contributions to retention from other solute-stationary phase interactions [31, 32, 37].

The relative importance of each type of interaction in the HSM to retention has also been investigated. Results have shown that the contribution of the hydrophobicity term in retention is the largest, as expected from the nature of RPLC separation [29, 30, 33]. The relative importance of other types of interaction is dependent on the particular solutes [29, 30, 33]. Given this fact, the Eq. 1.3 can be transferred into an approximate HSM (see Eq. 1.4),

$$\log \alpha \equiv \log\left(\frac{k}{k_{EB}}\right) \approx \eta' \mathbf{H} \qquad 1.4$$

Where only the hydrophobicity is considered as the primary contribution to the retention of compounds in RPLC. The HSM was originally developed for column selectivity, now it has been intensively used to select the equivalent or orthogonal columns for the separation of compounds of interest.

1.5.2 Column selectivity using the HSM

One of the most important applications of the HSM is column selectivity in RPLC. In the original work, the contribution of the various terms of Eq. 1.3 to column selectivity was investigated [29, 30, 33]. The hydrophobicity term is the least important contribution to changes in column selectivity, as the term of hydrophobicity is highly correlated with the retention [29, 30, 33]. Apart from hydrophobicity, each of the remaining terms of Eq. 1.3 can be important in determining column selectivity [31, 32, 37]. Until now, the HSM has been used in many fields with different applications in chromatography. However, it is worth pointing out that the most important focus of the HSM is mainly on the selection of "equivalent" columns to obtain the same separation through the use of a column comparison function based on the values of **H**, **S**, **A**, **B**, and **C** for two columns, or the selection of columns with very different selectivity for the development of orthogonal RPLC methods and two-dimensional (2D) separation [32, 127]. For example, an on-line column selectivity database for column comparison has been created based on the HSM. By using a calculated similarity factor (F_s) based on Eq. 1.5, a column then can be compared with other columns of interest, allowing equivalent or orthogonal columns to be selected rapidly [32, 87].

$$F_s = \sqrt{(w_H(H_1 - H_2))^2 + (w_S(S_1 - S_2))^2 + (w_A(A_1 - A_1))^2 + (w_B(B_1 - B_2))^2 + (w_{C_{2.8}}(c_{2.8_1} - c_{2.8_2}))^2} \quad 1.5$$

Where the **H**, **S**, **A**, **B**, and **C** are the column coefficients in the HSM for the column 1 and column 2, w is the weighting factor for the difference in each term of column coefficients for column 1 and column 2 [32, 87].

1.5.3 Retention prediction using the HSM

As mentioned previously, the HSM was not originally aimed to facilitate retention prediction, but a prediction accuracy of ±1-2% in *k* is claimed using this model suggesting that it has great potential for the purpose of retention prediction as well [29, 32]. Additionally, such a high level of accuracy indicates that the major contributors to the retention mechanism in RPLC have been considered. At this time, column coefficients for nearly 700 commercial C8 and C18 silica-based columns characterised using this HSM are available through an open-access database hosted by the U.S. Pharmacopeia Convention (USP) website [38, 72, 86]. Therefore, there is a unique opportunity for predicting the retention of potentially any given compound on all of the characterised columns, provided that the compound coefficients of the HSM are available.

1.6 Non-Targeted Metabolomics

1.6.1 Concept and background

Metabolomics studies can be targeted and non-targeted [128-130]. For targeted metabolomics, predefined metabolite-specific signals are often used for the quantitative measurement of a select group of known metabolites [129, 130]. In contrast, NTM involves high-throughput and comprehensive analysis of all the measurable analytes present in a given sample, thus it must be coupled to additional methods for subsequent interpretation by means of *in silico* or experimental routines [90, 130].

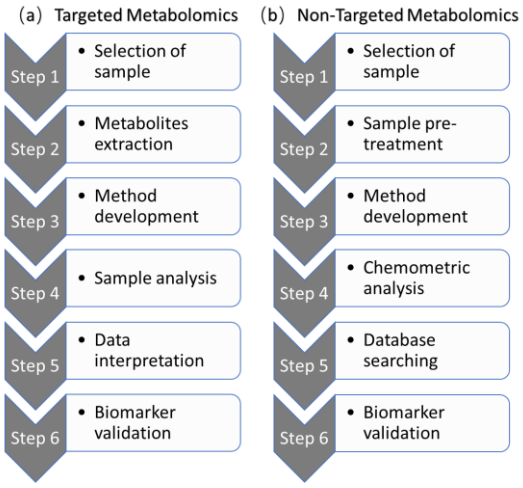


Figure 1.6. Schematic of (a) Targeted and (b) Non-Targeted Metabolomics workflow.

As Figure 1.6 shows, for targeted metabolomics, the identity of the metabolites of interest is the first step, followed by data acquisition where multiple simultaneous fragmentations are generated and analysed [130, 131]. The last step is data analysis and data interpretation. The process in NTM is quite different (Figure 1.6). Steps include sample selection and pre-treatment, sample analysis using NMR or MS coupled with separation techniques, data treatment and statistics, and finally, the identification and interpretation of metabolites using a commercial database or other analytical tools [132, 133].

In NTM, the ultimate goal is the identification of metabolites, allowing analytical data to be converted into meaningful biological knowledge [134-136]. However, a confident and unequivocal structure identification requires significant effort, which is multiplied dramatically in NTM where metabolites cover a diverse chemical space [130, 137]. Although integrated strategies like MS/NMR can provide much information for the identification of metabolites, authentic pure chemical standards of the metabolites of interest are still needed for unequivocal identification [28, 41, 134, 135]. At present, structure identification in NTM has remained costly, time-consuming, and frequently unsuccessful. This is because the process is complex and highly dependent on the robustness of the analytical platform and methods applied, as well as the databases and resources used for mass-based searching [132, 133, 138]. Metabolomics samples are typically complex and there are multiple interactions between metabolites in biological states, therefore, confident identification is still the bottleneck in NTM analysis.

1.6.2 Metabolite identification in NTM

LC-MS-based analysis is the most widely used analytical platform in NTM as it provides quantitative analysis with a combination of selectivity and sensitivity, and increases the possibility of identifying metabolites [28, 139]. The LC system for the separation of a sample can reduce ion suppression which is caused by co-eluting compounds, and isobaric interferences. In LC-MS-based NTM analysis, a target peak displaying metabolomics information is selected and the accurate mass of the eluted metabolite is identified [113, 140]. This accurate mass is then used to define potential molecular formulae corresponding to the chosen metabolite peak by searching on-line electronic resources, such as the Human Metabolome Database (HMDB), METLIN, LMSD, MassBank, RIKEN and PubChem [113, 140]. Based on the exact mass of the metabolite of interest, potential candidates can be found from the databases, and the returned matches then can be identified and confirmed using additional experimental data. The limitation of using mass searching is that in many cases, the candidate is not found, or alternatively a number of candidates are returned, which means that some false positives need to be removed [113, 136]. Those metabolites often behave

differently in biological systems but similarly in analytical platforms with respect to mass, which increases the probability of misidentification [28, 141, 142]. In this thesis, retention prediction QSRR models are explored that offer useful predictive ability for compounds having the same molecular weight, allowing false positives to be removed during the interpretation of structure identification in NTM.

1.7 Aims of project

The first goal of this project is to develop predictive QSRR strategy, enabling rapid method development in RPLC to speed up the *scoping* phase of chromatographic method development. The second goal is using the proposed QSRR strategy to remove false positives in NTM, enhancing the confidence and accuracy of metabolite identification. The research comprised a series of highly integrated research topics which cover the areas of structure-retention relationships for the RPLC mode, selection of the stationary phases based on predicted retention, co-elution prediction using the proposed strategy, and elimination of false positive identifications in NTM.

Aim 1: To build QSRR models based on molecular descriptors computed from chemical structures utilising a combination of GA and PLS, allowing the *scoping* phase of method development in RPLC to be accelerated for target analytes.

Aim 2: To develop strategies to improve the accuracy and predictivity of QSRR models using a combination of the concepts of structural similarity and chromatographic similarity. The employed strategies are to be based on different similarity filters combined with several approaches for compound classification to yield training sets for the construction of local QSRR models.

Aim 3: To establish QSRR models by predicting five solute coefficients (η', σ', β', α', and κ') in the HSM using the proposed QSRR strategy, allowing the retention of compounds of interest to be accurately predicted. The great advantage of QSRR modelling for the solute coefficients of analytes, is that nowadays over 700 columns in RPLC have been characterised by the HSM, and the parameters (**H**, **S**, **A**, **B**, and **C**) for all of those columns are available on an open database. So, there is a unique opportunity for predicting the retention of potentially any given solute on all of the characterised columns, albeit under only one isocratic condition, provided that the HSM solute coefficients can be accurately predicted for that solute.

Aim 4: To evaluate the application of the proposed QSRR strategy in eliminating false positives having the same mass, and improve the confidence of metabolite identification for LC-MS based Non-Target Metabolomics (NTM). The retention of analytes, which provides

information orthogonal to mass, makes it possible to eliminate analytes from the list of candidates that have the same exact mass based on the predicted retention.

1.8 References

1. Kazakevich, Y.V. and R. Lobrutto, *HPLC for pharmaceutical scientists*. 2007: John Wiley & Sons.
2. Rathore, A.S. and H. Winkle, *Quality by design for biopharmaceuticals.* Nature Biotechnology, 2009. **27**(1): p. 26-34.
3. Dong, M.W., *A Universal Reversed-Phase HPLC Method for Pharmaceutical Analysis.* LCGC North America, 2016. **34**(6): p. 408-419.
4. Visky, D., E. Haghedooren, P. Dehouck, Z. Kovács, K. Kóczián, B. Noszál, J. Hoogmartens, and E. Adams, *Facilitated column selection in pharmaceutical analyses using a simple column classification system.* Journal of Chromatography A, 2006. **1101**(1): p. 103-114.
5. Dong, M.W., *HPLC Column Standardization in Pharmaceutical Development: A Case Study.* LCGC Asia Pacific, 2016. **19**(3): p. 27-31.
6. Hewitt, E.F., P. Lukulay, and S. Galushko, *Implementation of a rapid and automated high performance liquid chromatography method development strategy for pharmaceutical drug candidates.* Journal of Chromatography A, 2006. **1107**(1-2): p. 79-87.
7. Krisko, R.M., K. McLaughlin, M.J. Koenigbauer, and C.E. Lunte, *Application of a column selection system and DryLab software for high-performance liquid chromatography method development.* Journal of Chromatography A, 2006. **1122**(1): p. 186-193.
8. Molnár, I., H.-J. Rieger, and R. Kormány, *Chromatography modelling in high performance liquid chromatography method development.* Chromatography Today, 2013. **6**(1): p. 3-8.
9. Borges, E.M., *How to select equivalent and complimentary reversed phase liquid chromatography columns from column characterization databases.* Analytica Chimica Acta, 2014. **807**: p. 143-152.
10. Héberger, K., *Quantitative structure–(chromatographic) retention relationships.* Journal of Chromatography A, 2007. **1158**(1-2): p. 273-305.
11. Desfontaine, V., D. Guillarme, E. Francotte, and L. Nováková, *Supercritical fluid chromatography in pharmaceutical analysis.* Journal of Pharmaceutical and Biomedical Analysis, 2015. **113**: p. 56-71.
12. Haddad, P.R. and P.E. Jackson, *Ion chromatography*. Vol. 46. 1990: Elsevier.
13. Dong, M.W., *Modern HPLC for practicing scientists*. 2006: John Wiley & Sons.
14. Claessens, H. and M. Van Straten, *Review on the chemical and thermal stability of stationary phases for reversed-phase liquid chromatography.* Journal of Chromatography A, 2004. **1060**(1): p. 23-41.
15. Kimata, K., K. Iwaguchi, S. Onishi, K. Jinno, R. Eksteen, K. Hosoya, M. Araki, and N. Tanaka, *Chromatographic characterization of silica C18 packing materials. Correlation between a preparation method and retention behavior of stationary phase.* Journal of Chromatographic Science, 1989. **27**(12): p. 721-728.
16. Khaledi, Morteza G., Joost K. Strasters, Andrew H. Rodgers, and Emelita D. Breyer. *Simultaneous enhancement of separation selectivity and solvent strength in reversed-phase liquid chromatography using micelles in hydro-organic solvents*. Analytical Chemistry 1990. **62**(2): p. 130-136.
17. Kromidas, S., *HPLC made to measure: a practical handbook for optimization*. 2008: John Wiley & Sons.
18. Robards, K., P.R. Haddad, and P.E. Jackson, *Principles and practice of modern chromatographic methods*. 1994: Academic Press.
19. Vogt, F.G. and A.S. Kord, *Development of quality-by-design analytical methods.* Journal of Pharmaceutical Sciences, 2011. **100**(3): p. 797-812.

20. Abate-Pella, D., D.M. Freund, Y. Ma, Y. Simón-Manso, J. Hollender, C.D. Broeckling, D.V. Huhman, O.V. Krokhin, D.R. Stoll, and A.D. Hegeman, *Retention projection enables accurate calculation of liquid chromatographic retention times across labs and methods.* Journal of Chromatography A, 2015. **1412**: p. 43-51.
21. Bolanča, T., Š. Ukić, M. Novak, and M. Rogošić, *Computer assisted method development in liquid chromatography.* Croatica Chemica Acta, 2014. **87**(2): p. 111-122.
22. García-Lavandeira, J., B. Losada, J. Martínez-Pontevedra, M. Lores, and R. Cela, *Computer-assisted method development in liquid chromatography–mass spectrometry: New proposals.* Journal of Chromatography A, 2008. **1208**(1): p. 116-125.
23. Cruz, E., M. Euerby, C. Johnson, and C. Hackett, *Chromatographic classification of commercially available reverse-phase HPLC columns.* Chromatographia, 1997. **44**(3): p. 151-161.
24. Karmarkar, S., R. Garber, Y. Genchanok, S. George, X. Yang, and R. Hammond, *Quality by design (QbD) based development of a stability indicating HPLC method for drug and impurities.* Journal of Chromatographic Science, 2011. **49**(6): p. 439-446.
25. Perisic-Janjic, N., R. Kaliszan, P. Wiczling, N. Milosevic, G. Uscumlic, and N. Banjac, *Reversed-phase TLC and HPLC retention data in correlation studies with in silico molecular descriptors and druglikeness properties of newly synthesized anticonvulsant succinimide derivatives.* Molecular Pharmaceutics, 2011. **8**(2): p. 555-563.
26. Horvath, C.G., B. Preiss, and S.R. Lipsky, *Fast liquid chromatography. Investigation of operating parameters and the separation of nucleotides on pellicular ion exchangers.* Analytical Chemistry, 1967. **39**(12): p. 1422-1428.
27. Kormány, R., J. Fekete, D. Guillarme, and S. Fekete, *Reliability of simulated robustness testing in fast liquid chromatography, using state-of-the-art column technology, instrumentation and modelling software.* Journal of Pharmaceutical and Biomedical Analysis, 2014. **89**: p. 67-75.
28. Lu, W., B.D. Bennett, and J.D. Rabinowitz, *Analytical strategies for LC–MS-based targeted metabolomics.* Journal of Chromatography B, 2008. **871**(2): p. 236-242.
29. Wilson, N., M. Nelson, J. Dolan, L. Snyder, R. Wolcott, and P. Carr, *Column selectivity in reversed-phase liquid chromatography: I. A general quantitative relationship.* Journal of Chromatography A, 2002. **961**(2): p. 171-193.
30. Wilson, N., M. Nelson, J. Dolan, L. Snyder, and P. Carr, *Column selectivity in reversed-phase liquid chromatography: II. Effect of a change in conditions.* Journal of Chromatography A, 2002. **961**(2): p. 195-215.
31. Snyder, L., A. Maule, A. Heebsh, R. Cuellar, S. Paulson, J. Carrano, L. Wrisley, C. Chan, N. Pearson, and J. Dolan, *A fast, convenient and rugged procedure for characterizing the selectivity of alkyl-silica columns.* Journal of Chromatography A, 2004. **1057**(1): p. 49-57.
32. Snyder, L., J. Dolan, and P. Carr, *The hydrophobic-subtraction model of reversed-phase column selectivity.* Journal of Chromatography A, 2004. **1060**(1): p. 77-116.
33. Wilson, N., J. Dolan, L. Snyder, P. Carr, and L.C. Sander, *Column selectivity in reversed-phase liquid chromatography: III. The physico-chemical basis of selectivity.* Journal of Chromatography A, 2002. **961**(2): p. 217-236.
34. Carr, P., J. Dolan, U. Neue, and L. Snyder, *Contributions to reversed-phase column selectivity. I. Steric interaction.* Journal of Chromatography A, 2011. **1218**(13): p. 1724-1742.
35. Marchand, D., P. Carr, D.V. McCalley, U. Neue, J. Dolan, and L. Snyder, *Contributions to reversed-phase column selectivity. II. Cation exchange.* Journal of Chromatography A, 2011. **1218**(40): p. 7110-7129.
36. Carr, P., J. Dolan, J. Dorsey, L. Snyder, and J. Kirkland, *Contributions to reversed-phase column selectivity: III. Column hydrogen-bond basicity.* Journal of Chromatography A, 2015. **1395**: p. 57-64.
37. Dolan, J., A. Maule, D. Bingley, L. Wrisley, C. Chan, M. Angod, C. Lunte, R. Krisko, J. Winston, and B. Homeier, *Choosing an equivalent replacement column for a reversed-*

phase liquid chromatographic assay procedure. Journal of Chromatography A, 2004. **1057**(1): p. 59-74.
38. Talebi, M., S.H. Park, M. Taraji, Y. Wen, R.I. Amos, P.R. Haddad, R. Shellie, R. Szucs, C. Pohl, and J.W. Dolan, *Retention time prediction based on molecular structure in pharmaceutical method development: A perspective.* LCGC North America, 2016. **34**(8): p. 550-558.
39. De Beer, M., F.d. Lynen, K. Chen, P. Ferguson, M. Hanna-Brown, and P. Sandra, *Stationary-phase optimized selectivity liquid chromatography: development of a linear gradient prediction algorithm.* Analytical Chemistry, 2010. **82**(5): p. 1733-1743.
40. Kormány, R., I. Molnár, J. Fekete, D. Guillarme, and S. Fekete, *Robust UHPLC separation method development for multi-API product containing amlodipine and bisoprolol: the impact of column selection.* Chromatographia, 2014. **77**(17-18): p. 1119-1127.
41. Meyer, M.R. and H.H. Maurer, *Current applications of high-resolution mass spectrometry in drug metabolism studies.* Analytical and Bioanalytical Chemistry, 2012. **403**(5): p. 1221-1231.
42. Ichida, A., T. Shibata, I. Okamoto, Y. Yuki, H. Namikoshi, and Y. Toga, *Resolution of enantiomers by HPLC on cellulose derivatives.* Chromatographia, 1984. **19**(1): p. 280-284.
43. Chan, F., L. Yeung, R. LoBrutto, and Y. Kazakevich, *Characterization of phenyl-type HPLC adsorbents*. Journal of Chromatography A, 2005. **1069**(2): p. 217-224.
44. Lee, D.P., *Reversed-phase HPLC from pH 1 to 13.* Journal of Chromatographic Science, 1982. **20**(5): p. 203-208.
45. Kazakevich, Y. and H. McNair, *Thermodynamic definition of HPLC dead volume*. Journal of Chromatographic Science, 1993. **31**(8): p. 317-322.
46. Plzák, Z., F.P. Dousek, and J. Jansta, *New carbon adsorbent for high-performance liquid chromatography*. Journal of Chromatography A, 1978. **147**: p. 137-142.
47. Hosoya, K., K. Yoshizako, N. Tanaka, K. Kimata, T. Araki, and J. Haginaka, *Uniform-size macroporous polymer-based stationary phase for HPLC prepared through molecular imprinting technique.* Chemistry Letters, 1994. **23**(8): p. 1437-1438.
48. Urban, J., S. Eeltink, P. Jandera, and P.J. Schoenmakers, *Characterization of polymer-based monolithic capillary columns by inverse size-exclusion chromatography and mercury-intrusion porosimetry*. Journal of Chromatography A, 2008. **1182**(2): p. 161-168.
49. Walter, T.H., P. Iraneta, and M. Capparella, *Mechanism of retention loss when C8 and C18 HPLC columns are used with highly aqueous mobile phases.* Journal of Chromatography A, 2005. **1075**(1-2): p. 177-183.
50. Kempe, M., *Antibody-mimicking polymers as chiral stationary phases in HPLC.* Analytical Chemistry, 1996. **68**(11): p. 1948-1953.
51. O'Gara, J.E., B.A. Alden, T.H. Walter, J.S. Petersen, C.L. Niederlaender, and U.D. Neue, *Simple preparation of a C8 HPLC stationary phase with an internal polar functional group.* Analytical Chemistry, 1995. **67**(20): p. 3809-3813.
52. Tanaka, N., T. Ebata, K. Hashizume, K. Hosoya, and M. Araki, *Polymer-based packing materials with alkyl backbones for reversed-phase liquid chromatography: Performance and retention selectivity*. Journal of Chromatography A, 1989. **475**(2): p. 195-208.
53. Nahum, A. and C. Horváth, *Surface silanols in silica-bonded hydrocarbonaceous stationary phases: I. Dual retention mechanism in reversed-phase chromatography.* Journal of Chromatography A, 1981. **203**: p. 53-63.
54. Layne, J., *Characterization and comparison of the chromatographic performance of conventional, polar-embedded, and polar-endcapped reversed-phase liquid chromatography stationary phases.* Journal of Chromatography A, 2002. **957**(2): p. 149-164.
55. Przybyciel, M. and R.E. Majors, *Phase collapse in reversed-phase liquid chromatography.* LCGC North America, 2002. **20**(6): p. 516-523.

56. Kirkland, J., J. Henderson, J. Martosella, B. Bidlingmeyer, J. Vasta-Russell, and J. Adams Jr, *A highly stable alkyl-amide silica-based column packing for reversed-phase HPLC of polar and ionizable compounds.* LC GC: Magazine of Liquid and Gas Chromatography, 1999. **17**(7): p. 634-639.
57. Neue, U.D., K. Van Tran, P.C. Iraneta, and B.A. Alden, *Characterization of HPLC packings.* Journal of Separation Science, 2003. **26**(3-4): p. 174-186.
58. Okamoto, Y. and T. Ikai, *Chiral HPLC for efficient resolution of enantiomers.* Chemical Society Reviews, 2008. **37**(12): p. 2593-2608.
59. Polyakova, Y., Y.M. Koo, and K.H. Row, *Application of ionic liquids as mobile phase modifier in HPLC.* Biotechnology and Bioprocess Engineering, 2006. **11**(1): p. 1.
60. Van der Wal, S., *Low viscosity organic modifiers in reversed-phase HPLC.* Chromatographia, 1985. **20**(5): p. 274-278.
61. Armstrong, D. W., S. Chen, C. Chang, and S. Chang. *A new approach for the direct resolution of racemic beta adrenergic blocking agents by HPLC*. Journal of Liquid Chromatography 1992. **15**(3): P. 545-556.
62. McCalley, D.V., *Selection of suitable stationary phases and optimum conditions for their application in the separation of basic compounds by reversed-phase HPLC.* Journal of Separation Science, 2003. **26**(3-4): p. 187-200.
63. Lu, H., M.A. Schmidt, and K.F. Jensen, *Photochemical reactions and on-line UV detection in microfabricated reactors.* Lab on a Chip, 2001. **1**(1): p. 22-28.
64. Snyder, L., M. Quarry, and J. Glajch, *Solvent-strength selectivity in reversed-phase HPLC.* Chromatographia, 1987. **24**(1): p. 33-44.
65. Feibush, B., A. Figueroa, R. Charles, K.D. Onan, P. Feibush, and B.L. Karger, *Chiral separation of heterocyclic drugs by HPLC: solute-stationary phase base-pair interactions.* Journal of the American Chemical Society, 1986. **108**(12): p. 3310-3318.
66. Berthod, A., S.S. Chang, J.P. Kullman, and D.W. Armstrong, *Practice and mechanism of HPLC oligosaccharide separation with a cyclodextrin bonded phase.* Talanta, 1998. **47**(4): p. 1001-1012.
67. Park, S.H., P.R. Haddad, M. Talebi, E. Tyteca, R.I. Amos, R. Szucs, J.W. Dolan, and C.A. Pohl, *Retention prediction of low molecular weight anions in ion chromatography based on quantitative structure-retention relationships applied to the linear solvent strength model.* Journal of Chromatography A, 2017. **1486**: p. 68-75.
68. Jeong, L.N., R. Sajulga, S.G. Forte, D.R. Stoll, and S.C. Rutan, *Simulation of elution profiles in liquid chromatography I: Gradient elution conditions, and with mismatched injection and mobile phase solvents.* Journal of Chromatography A, 2016. **1457**: p. 41-49.
69. Bosch, E., P. Bou, H. Allemann, and M. Rosés, *Retention of Ionizable Compounds on HPLC. pH Scale in Methanol− Water and the p K and pH Values of Buffers.* Analytical Chemistry, 1996. **68**(20): p. 3651-3657.
70. Dolan, J.W., *Temperature selectivity in reversed-phase high performance liquid chromatography.* Journal of Chromatography A, 2002. **965**(1-2): p. 195-205.
71. Talebi, M., G. Schuster, R.A. Shellie, R. Szucs, and P.R. Haddad, *Performance comparison of partial least squares-related variable selection methods for quantitative structure retention relationships modelling of retention times in reversed-phase liquid chromatography.* Journal of Chromatography A, 2015. **1424**: p. 69-76.
72. Wen, Y., M. Talebi, R.I. Amos, R. Szucs, J.W. Dolan, C.A. Pohl, and P.R. Haddad, *Retention prediction in reversed phase high performance liquid chromatography using quantitative structure-retention relationships applied to the Hydrophobic Subtraction Model.* Journal of Chromatography A, 2018. **1541**: p. 1-11.
73. Taraji, M., P.R. Haddad, R.I. Amos, M. Talebi, R. Szucs, J.W. Dolan, and C.A. Pohl, *Prediction of retention in hydrophilic interaction liquid chromatography using solute molecular descriptors based on chemical structures.* Journal of Chromatography A, 2017. **1486**: p. 59-67.
74. Park, S.H., P.R. Haddad, R.I. Amos, M. Talebi, R. Szucs, C.A. Pohl, and J.W. Dolan, *Towards a chromatographic similarity index to establish localised Quantitative Structure-*

Retention Relationships for retention prediction. III Combination of Tanimoto similarity index, logP, and retention factor ratio to identify optimal analyte training sets for ion chromatography. Journal of Chromatography A, 2017. **1520**: p. 107-116.
75. Tyteca, E., A. Liekens, D. Clicq, A. Fanigliulo, B. Debrus, S. Rudaz, D. Guillarme, and G. Desmet, *Predictive elution window stretching and shifting as a generic search strategy for automated method development for liquid chromatography*. Analytical Chemistry, 2012. **84**(18): p. 7823-7830.
76. Larsen, E.H., *Method optimization and quality assurance in speciation analysis using high performance liquid chromatography with detection by inductively coupled plasma mass spectrometry*. Spectrochimica Acta Part B: Atomic Spectroscopy, 1998. **53**(2): p. 253-265.
77. Schmidt, A.H., M. Stanic, and I. Molnár, *In silico robustness testing of a compendial HPLC purity method by using of a multidimensional design space build by chromatography modeling—Case study pramipexole.* Journal of Pharmaceutical and Biomedical Analysis, 2014. **91**: p. 97-107.
78. Taylor, T., *Relating Analyte Properties to HPLC Column Selectivity–On the Road to Nirvana.* LCGC North America, 2017.
79. Torres-Lapasió, J. and M. García-Álvarez-Coque, *Levels in the interpretive optimisation of selectivity in high-performance liquid chromatography: A magical mystery tour.* Journal of Chromatography A, 2006. **1120**(1-2): p. 308-321.
80. Andrić, F. and K. Héberger, *How to compare separation selectivity of high-performance liquid chromatographic columns properly?* Journal of Chromatography A, 2017. **1488**: p. 45-56.
81. Coffman, J.L., J.F. Kramarczyk, and B.D. Kelley, *High-throughput screening of chromatographic separations: I. Method development and column modeling.* Biotechnology and Bioengineering, 2008. **100**(4): p. 605-618.
82. Tyteca, E., S.H. Park, R.A. Shellie, P.R. Haddad, and G. Desmet, *Computer-assisted multi-segment gradient optimization in ion chromatography.* Journal of Chromatography A, 2015. **1381**: p. 101-109.
83. Zamora, I., F. Fontaine, B. Serra, and G. Plasencia, *High-throughput, computer assisted, specific MetID. A revolution for drug discovery*. Drug Discovery Today: Technologies, 2013. **10**(1): p. 199-205.
84. Nolvachai, Y., C. Kulsing, and P.J. Marriott, *In Silico Modeling of Hundred Thousand Experiments for Effective Selection of Ionic Liquid Phase Combinations in Comprehensive Two-Dimensional Gas Chromatography.* Analytical Chemistry , 2016. **88**(4): p. 2125-2131.
85. Dearden, J.C., *In silico prediction of drug toxicity*. Journal of Computer-Aided Molecular Design, 2003. **17**(2): p. 119-127.
86. *The United States Pharmacopeia Convention*, http://www.usp.org/resources/pqri-approach-column-equiv-tool.
87. *Dwight Stoll, Paul Boswell*, http://www.hplccolumns.org/. Available from: http://www.hplccolumns.org.
88. Karmarkar, S., R. Garber, Y. Genchanok, S. George, X. Yang, and R. Hammond, *Quality by design (QbD) based development of a stability indicating HPLC method for drug and impurities.* Journal of Chromatographic Science, 2011. **49**(6): p. 439-446.
89. Kaliszan, R., *Quantitative structure-retention relationships.* Analytical Chemistry, 1992. **64**(11): p. 619A-631A.
90. Goryński, K., B. Bojko, A. Nowaczyk, A. Buciński, J. Pawliszyn, and R. Kaliszan, *Quantitative structure–retention relationships models for prediction of high performance liquid chromatography retention time of small molecules: endogenous metabolites and banned compounds.* Analytica Chimica Acta, 2013. **797**: p. 13-19.
91. Tyteca, E., M. Talebi, R. Amos, S.H. Park, M. Taraji, Y. Wen, R. Szucs, C.A. Pohl, J.W. Dolan, and P.R. Haddad, *Towards a chromatographic similarity index to establish localized quantitative structure-retention models for retention prediction: use of retention factor ratio.* Journal of Chromatography A, 2017. **1486**: p. 50-58.

92. Taraji, M., P.R. Haddad, R.I. Amos, M. Talebi, R. Szucs, J.W. Dolan, and C.A. Pohl, *Use of dual-filtering to create training sets leading to improved accuracy in quantitative structure-retention relationships modelling for hydrophilic interaction liquid chromatographic systems.* Journal of Chromatography A, 2017. **1507**: p. 53-62.
93. Kaliszan, R., *Quantitative structure-retention relationships applied to reversed-phase high-performance liquid chromatography.* Journal of Chromatography A, 1993. **656**(1-2): p. 417-435.
94. Put, R. and Y. Vander Heyden, *Review on modelling aspects in reversed-phase liquid chromatographic quantitative structure–retention relationships.* Analytica Chimica Acta, 2007. **602**(2): p. 164-172.
95. *R. Todeschini, V. Consonni, Handbook of Molecular Descriptors*. Wiley-WCH, Weinheim, 2000.
96. Todeschini, R. and V. Consonni, *Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references*. Vol. 41. 2009: John Wiley & Sons.
97. Guba, W. and G. Cruciani, *Molecular field-derived descriptors for the multivariate modeling of pharmacokinetic data*, in *Molecular modeling and prediction of bioactivity*. 2000, Springer. p. 89-94.
98. *in, Talete srl, Dragon 6.0 for Windows (Software For Molecular Descriptor Calculations);* http://www.talete.mi.it/ Talete, Milano, Italy.
99. Cruciani, G., P. Crivori, P.-A. Carrupt, and B. Testa, *Molecular fields in quantitative structure–permeation relationships: the VolSurf approach.* Journal of Molecular Structure: THEOCHEM, 2000. **503**(1): p. 17-30.
100. Cruciani, G., M. Pastor, and W. Guba, *VolSurf: a new tool for the pharmacokinetic optimization of lead compounds.* European Journal of Pharmaceutical Sciences, 2000. **11**: p. S29-S39.
101. Goodarzi, M., R. Jensen, and Y. Vander Heyden, *QSRR modeling for diverse drugs using different feature selection methods coupled with linear and nonlinear regressions.* Journal of Chromatography B, 2012. **910**: p. 84-94.
102. Ghasemi, J. and S. Saaidpour, *QSRR prediction of the chromatographic retention behavior of painkiller drugs.* Journal of Chromatographic Science, 2009. **47**(2): p. 156-163.
103. Todeschini, R. and V. Consonni, *Handbook of molecular descriptors*. Vol. 11. 2008: John Wiley & Sons.
104. Almeida, T.M., A. Leitão, M.L.C. Montanari, and C.A. Montanari, *The Molecular Retention Mechanism in Reversed-Phase Liquid Chromatography of Meso-ionic Compounds by Quantitative Structure–Retention Relationships (QSRR).* Chemistry & Biodiversity, 2005. **2**(12): p. 1691-1700.
105. Cruciani, G., M. Pastor, and S. Clementi, *Handling information from 3D grid maps for QSAR studies*, in *Molecular modeling and prediction of bioactivity*. 2000, Springer. p. 73-81.
106. Stephens, P., F. Devlin, C. Chabalowski, and M.J. Frisch, *Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields.* Journal of Physical Chemistry, 1994. **98**(45): p. 11623-11627.
107. Clementi, S., G. Cruciani, P. Fifi, D. Riganelli, R. Valigi, and G. Musumarra, *A new set of principal properties for heteroaromatics obtained by GRID.* Molecular Informatics, 1996. **15**(2): p. 108-120.
108. *VolSurf+ 1.0.7.1 software (Molecular Discovery Ltd., Hertfordshire, UK),* http://www.moldiscovery.com/software/vsplus/.
109. Leardi, R., *Application of genetic algorithm-PLS for feature selection in spectral data sets.* Journal of Chemometrics, 2000. **14**(5-6): p. 643-655.
110. Varmuza, K. and P. Filzmoser, *Introduction to multivariate statistical analysis in chemometrics*. 2016: CRC press.

111. Leardi, R. and A.L. Gonzalez, *Genetic algorithms applied to feature selection in PLS regression: how and when to use them.* Chemometrics and Intelligent Laboratory Systems , 1998. **41**(2): p. 195-207.
112. John, H., *Holland, Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. 1992, MIT Press, Cambridge, MA.
113. Hall, L.M., D.W. Hill, L.C. Menikarachchi, M.-H. Chen, L.H. Hall, and D.F. Grant, *Optimizing artificial neural network models for metabolomics and systems biology: an example using HPLC retention index data.* Bioanalysis, 2015. **7**(8): p. 939-955.
114. Riahi, S., M.R. Ganjali, E. Pourbasheer, and P. Norouzi, *QSRR study of GC retention indices of essential-oil compounds by multiple linear regression with a genetic algorithm.* Chromatographia, 2008. **67**(11-12): p. 917-922.
115. Varmuza, K., P. Filzmoser, and M. Dehmer, *Multivariate linear QSPR/QSAR models: Rigorous evaluation of variable selection for PLS.* Computational and Structural Biotechnology Journal, 2013. **5**(6): p. e201302007.
116. Vainio, M.J. and M.S. Johnson, *Generating conformer ensembles using a multiobjective genetic algorithm.* Journal of Chemical Information and Modeling, 2007. **47**(6): p. 2462-2474.
117. Žuvela, P., J.J. Liu, K. Macur, and T. Baczek, *Molecular descriptor subset selection in theoretical peptide quantitative structure–retention relationship model development using nature-inspired optimization algorithms.* Analytical Chemistry, 2015. **87**(19): p. 9876-9883.
118. Wiklund, S., D. Nilsson, L. Eriksson, M. Sjöström, S. Wold, and K. Faber, *A randomization test for PLS component selection.* Journal of Chemometrics, 2007. **21**(10-11): p. 427-439.
119. Wang, C., M.J. Skibic, R.E. Higgs, I.A. Watson, H. Bui, J. Wang, and J.M. Cintron, *Evaluating the performances of quantitative structure-retention relationship models with different sets of molecular descriptors and databases for high-performance liquid chromatography predictions.* Journal of Chromatography A, 2009. **1216**(25): p. 5030-5038.
120. Muteki, K., J.E. Morgado, G.L. Reid, J. Wang, G. Xue, F.W. Riley, J.W. Harwood, D.T. Fortin, and I.J. Miller, *Quantitative structure retention relationship models in an analytical Quality by Design framework: simultaneously accounting for compound properties, mobile-phase conditions, and stationary-phase properties.* Industrial & Engineering Chemistry Research, 2013. **52**(35): p. 12269-12284.
121. Johnson, M.A. and G.M. Maggiora, *Concepts and applications of molecular similarity*. 1990: Wiley.
122. Sheridan, R.P., B.P. Feuston, V.N. Maiorov, and S.K. Kearsley, *Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR.* Journal of Chemical Information and Computer Sciences, 2004. **44**(6): p. 1912-1928.
123. Yuan, H., Y. Wang, and Y. Cheng, *Local and global quantitative structure− activity relationship modeling and prediction for the baseline toxicity.* Journal of Chemical Information and Modeling, 2007. **47**(1): p. 159-169.
124. Bergström, C.A., C.M. Wassvik, U. Norinder, K. Luthman, and P. Artursson, *Global and local computational models for aqueous solubility prediction of drug-like molecules.* Journal of Chemical Information and Computer Sciences, 2004. **44**(4): p. 1477-1488.
125. Willett, P., *Similarity methods in chemoinformatics.* Annual Review of Information Science and Technology, 2009. **43**: p. 3-71.
126. Willett, P., *Similarity-based virtual screening using 2D fingerprints.* Drug Discovery Today, 2006. **11**(23-24): p. 1046-1053.
127. Græsbøll, R., N.J. Nielsen, and J.H. Christensen, *Using the hydrophobic subtraction model to choose orthogonal columns for online comprehensive two-dimensional liquid chromatography*. Journal of Chromatography A, 2014. **1326**: p. 39-46.

128. Aicheler, F., J. Li, M. Hoene, R. Lehmann, G. Xu, and O. Kohlbacher, *Retention time prediction improves identification in nontargeted lipidomics approaches.* Analytical Chemistry, 2015. **87**(15): p. 7698-7704.
129. Griffiths, W.J., T. Koal, Y. Wang, M. Kohl, D.P. Enot, and H.P. Deigner, *Targeted metabolomics for biomarker discovery.* Angewandte Chemie International Edition, 2010. **49**(32): p. 5426-5445.
130. Naz, S., M. Vallejo, A. García, and C. Barbas, *Method validation strategies involved in non-targeted metabolomics.* Journal of Chromatography A, 2014. **1353**: p. 99-105.
131. Christians, U., J. Klawitter, A. Hornberger, and J. Klawitter, *How unbiased is non-targeted metabolomics and is targeted pathway screening the solution?* Current Pharmaceutical Biotechnology, 2011. **12**(7): p. 1053-1066.
132. Riedl, J., S. Esslinger, and C. Fauhl-Hassek, *Review of validation and reporting of non-targeted fingerprinting approaches for food authentication.* Analytica Chimica Acta, 2015. **885**: p. 17-32.
133. Mairinger, T., T.J. Causon, and S. Hann, *The potential of ion mobility–mass spectrometry for non-targeted metabolomics.* Current Opinion in Chemical Biology, 2018. **42**: p. 9-15.
134. Putri, S.P., S. Yamamoto, H. Tsugawa, and E. Fukusaki, *Current metabolomics: technological advances.* Journal of Bioscience and Bioengineering, 2013. **116**(1): p. 9-16.
135. Fuhrer, T. and N. Zamboni, *High-throughput discovery metabolomics.* Current Opinion in Biotechnology, 2015. **31**: p. 73-78.
136. Creek, D.J., W.B. Dunn, O. Fiehn, J.L. Griffin, R.D. Hall, Z. Lei, R. Mistrik, S. Neumann, E.L. Schymanski, and L.W. Sumner, *Metabolite identification: are you sure? And how do your peers gauge your confidence?* Metabolomics, 2014. **10**(3): p. 350-353.
137. Liu, Y., K. Smirnov, M. Lucio, R.D. Gougeon, H. Alexandre, and P. Schmitt-Kopplin, *MetICA: independent component analysis for high-resolution mass-spectrometry based non-targeted metabolomics.* BMC Bioinformatics, 2016. **17**(1): p. 114.
138. Díaz, R., H. Gallart-Ayala, J.V. Sancho, O. Nunez, T. Zamora, C.P. Martins, F. Hernández, S. Hernández-Cassou, J. Saurina, and A. Checa, *Told through the wine: A liquid chromatography–mass spectrometry interplatform comparison reveals the influence of the global approach on the final annotated metabolites in non-targeted metabolomics.* Journal of Chromatography A, 2016. **1433**: p. 90-97.
139. DeHaven, C.D., A.M. Evans, H. Dai, and K.A. Lawton, *Organization of GC/MS and LC/MS metabolomics data into chemical libraries.* Journal of Cheminformatics, 2010. **2**(1): p. 9.
140. Wei, R., G. Li, and A.B. Seymour, *High-throughput and multiplexed LC/MS/MRM method for targeted metabolomics.* Analytical Chemistry, 2010. **82**(13): p. 5527-5533.
141. Ganna, A., S. Salihovic, J. Sundström, C.D. Broeckling, Å.K. Hedman, P.K. Magnusson, N.L. Pedersen, A. Larsson, A. Siegbahn, and M. Zilmer, *Large-scale metabolomic profiling identifies novel biomarkers for incident coronary heart disease.* PLoS Genetics, 2014. **10**(12): p. e1004801.
142. Bueno, M.J.M., F.J. Díaz-Galiano, Ł. Rajski, V. Cutillas, and A.R. Fernández-Alba, *A non-targeted metabolomic approach to identify food markers to support discrimination between organic and conventional tomato crops.* Journal of Chromatography A, 2018.

2 Experimental Section and Data Collection

2.1 Databases

2.1.1 Databases for retention prediction

Three RPLC databases were employed in the present study. The first database (Dataset 1) was originally used by Wilson *et al.* to derive the Hydrophobic Subtraction Model (HSM), and consists of ten columns and 90 compounds [1]. The characteristics of the ten C18 Columns used in Dataset 1 are listed in Table 2.1 (numbered from 1 to 10), the compounds in Dataset 1 are listed in Table 2.2. The second (Dataset 2), originally reported by Tan *et al*., was used by Wilson *et al*. in their work and consists of five columns and 87 compounds (Table 2.3) [2]. The five columns for Dataset 2 are: Zorbax SB-C18 (column number 11), Zorbax Rx-C18 (column number 12), Hypersil C18 (column number 13), Hypersil C8 (column number 14) and Zorbax C8 (column number 15). The third database (Dataset 3) is an open-access database from Boswell Research group [3], containing the retention data of 112 compounds on a Zorbax Eclipse Plus C18 column (column number 16), the compounds are listed in Table 2.4. The retention factors of the compounds in each dataset were also extracted and listed below (Tables 2.5 to 2.7). The retention of compounds numbered from 68 to 90 in Dataset 1 was only performed on nine columns (column number 1-8, and column number 10), therefore the retention of these compounds on column number 9 was unavailable (Table 2.5).

Table 2.1. Characteristics of ten C18 columns used in dataset 1 in present study; 5-μm particles, 150×4.6 mm column dimensions

Column	Abbr.	Surface area (m^2/g)	Pore diameter (nm)	% C	μmol/m^2	Metal	
						Fe	AI
GL Inertsil ODS-3	C1	436	9.5	14.7	1.74	2.8	<0.5
Waters Symmetry C18	C2	343	9	19.7	3.13	<10	<10
HP Zorbax SB C18	C3	186	8	10.4	2.08	<1	<1
HP Zorbax SB C18	C4	188	8	9.2	1.79	<1	<1
HP Zorbax SB-300 C18	C5	52	30	3.25	2.09	<1	<1
HP Eclipse XDB-C18	C6	186	8	10.7	3	<1	<1
YMC Pack Pro C18	C7	322	12.5	15.5	2.51	<10	<10
YMC Pack Pro C18	C8	321	12.5	16.3	2.68	<10	<10
YMC Pack Pro C18	C9	322	12.5	17	2.82	<10	<10
Supelco Discovery C18	C10	190-220	17-20	12.5	3.12	<20	<1

Table 2.2. Compounds used (Dataset 1) in the present study

ID	Name	ID	Name
1	benzene	46	amitriptyline
2	toluene	47	diphenhydramine
3	ethylbenzene	48	D,L-propanolol
4	p-xylene	49	nortriptyline
5	propylbenzene	50	prolintane
6	butylbenzene	51	4-n-pentylaniline
7	naphthalene	52	4-n-hexylaniline
8	p-chlorotoluene	53	4-n-heptylaniline
9	Dichlorobenzene	54	N-ethylaniline
10	benzotrichloride	55	2-phenylpyridine
11	bromobenzene	56	diclofenac
12	1-nitropropane	57	mefenamic
13	nitrobenzene	58	ketoprofen
14	p-nitrotoluene	59	diflunisal
15	p-nitrobenzylchloride	60	4-n-butylbenzoic acid
16	N-benzylformamide	61	4-n-pentylbenzoic acid
17	anisole	62	4-n-hexylbenzoic acid
18	benzylalcohol	63	3-cyanobenzoic acid
19	3-phenylpropanol	64	2-nitrobenzoic acid
20	5-phenylpentanol	65	3-nitrobenzoic acid
21	phenol	66	2,6-dimethylbenzoic acid
22	p-chlorophenol	67	2-fluorobenzoic acid
23	2,3-dihydroxynaphthalene	68	1,2-dinitrobenzene
24	1,3-dihydroxynaphthalene	69	1,3-dinitrobenzene
25	eugenol	70	nitrocyclohexane
26	danthron	71	biphenyl
27	n-propylformate	72	2-nitrobiphenyl
28	methylbenzoate	73	3-nitrobiphenyl
29	benzonitrile	74	2-biphenylmethanol
30	Coumarin	75	2,2'-biphenol
31	acetophenone	76	4,4'-biphenol
32	benzophenone	77	diphenylbutyrolactone
33	cis-chalcone	78	fluorescamine
34	trans-chalcone	79	camphorquinone
35	cis-4-nitrochalcone	80	ferrocene
36	trans-4-nitrochalcone	81	N,N-diethylacetamide
37	cis-4-methoxychalcone	82	3-nitrophenol
38	trans-4-methoxychalcone	83	4-nitrophenol
39	prednisone	84	2,4-dinitrophenol
40	hydrocortisone	85	2-5-dinitrophenol
41	mephenytoin	86	picric acid
42	oxazepam	87	fisetin
43	flunitrazepam	88	biochanin A
44	5,5-dimethylhydantoin	89	4-phenylpyridine
45	N,N-dimethylacetamide	90	N-butylaniline

Table 2.3. Compounds used (Dataset 2) in the present study

ID	Name	ID	Name
1	1-butanol	45	3-phenyl propanol
2	1-hexanol	46	benzaldehyde
3	1-octanol	47	N-benzyl formamide
4	2-propanol	48	methyl benzoate
5	cyclohexanol	49	ethyl benzoate
6	1-butanal	50	anisole
7	1-hexanal	51	acetophenone
8	1-heptanal	52	propiophenone
9	1-octanal	53	benzophenone
10	N,N-dimethyl formamide	54	benzonitrile
11	N,N-diethyl formamide	55	m-toluenitrile
12	N,N-dibutyl formamide	56	benzyl cyanide
13	N,N-dimethyl acetamide	57	nitrobenzene
14	N,N-diethyl acetamide	58	m-nitrotoluene
15	n-propyl formate	59	o-nitrotoluene
16	n-butyl acetate	60	p-nitrotoluene
17	n-amyl acetate	61	p-nitrobenzyl bromide
18	n-hexyl acetate	62	p-nitrobenzyl chloride
19	ethyl propionate	63	fluorobenzene
20	ethyl butyrate	64	chlorobenzene
21	ethyl ether	65	bromobenzene
22	n-propyl ether	66	lodobenzene
23	n-butyl ether	67	benzyl bromide
24	dioxane	68	p-chlorotoluene
25	acetone	69	p-bromotoluene
26	2-butanone	70	p-dichlorobenzene
27	2-hexanone	71	benzene
28	2-heptanone	72	toluene
29	2-nonanone	73	ethylbenzene
30	cyclopentanone	74	n-propylbenzene
31	n-propionitrile	75	n-butylbenzene
32	n-valeronitrile	76	tert.-butylbenzene
33	n-hexanitrile	77	p-xylene
34	n-hexyl cyanide	78	mesitylene
35	n-heptyl cyanide	79	biphenyl
36	n-octyl cyanide	80	naphthalene
37	n-nitropropane	81	anthracene
38	n-nitrobutane	82	phenol
39	n-nitropentane	83	m-cresol
40	methylene chloride	84	p-cresol
41	chloroform	85	o-cresol
42	dibromomethane	86	p-ethylphenol
43	benzyl alcohol	87	p-chlorophenol
44	2-phenyl ethanol		

Table 2.4. Compounds used (Dataset 3) in the present study

ID	Name
1	N-[1-hydroxy-3-(morpholin-4-yl)-1-phenylpropan-2-yl]decanamide
2	2-(6-chloro-1H-indol-3-yl)acetic acid
3	2-(2-amino-3-methoxyphenyl)-4H-chromen-4-one
4	1-(2,3-dihydro-1,4-benzodioxine-6-carbonyl)piperidine
5	2-(3,4-dimethoxyphenyl)-5,6,7-trimethoxy-4H-chromen-4-one
6	4-[(R)-[(2S,5R)-2,5-dimethyl-4-(prop-2-en-1-yl)piperazin-1-yl](3-methoxyphenyl)methyl]-N,N-diethylbenzamide
7	6-methoxy-2-(4-methoxyphenyl)-4H-chromen-4-one
8	1H-indole-6-carbonitrile
9	2-(7-chloro-1H-indol-3-yl)acetic acid
10	7-methyl-1H-indole
11	(1S,2R,10S,11S,14S,15S,17R)-14-acetyl-2,15-dimethyl-5-oxotetracyclo[8.7.0.0^{2},^{7}.0^{11},^{15}]heptadec-6-en-17-yl acetate
12	(1R,2S,4R,6S,7S,10S,11R,14S)-6-acetyl-7,11-dimethyl-5-oxapentacyclo[8.8.0.0^{2},^{7}.0^{4},^{6}.0^{11},^{16}]octadec-16-en-14-yl acetate
13	1-(4,6-dimethoxypyrimidin-2-yl)-3-[methanesulfonyl(methyl)sulfamoyl]urea
14	1-(4-methoxybenzoyl)pyrrolidin-2-one
15	5,7-dihydroxy-2-(4-hydroxyphenyl)-4H-chromen-4-one
16	1-{[2-(2,4-dichlorophenyl)-1,3-dioxolan-2-yl]methyl}-1H-1,2,4-triazole
17	(1R,2S,10S,11S,13S,14R,15S,17S)-1-chloro-14,17-dihydroxy-14-(2-hydroxyacetyl)-2,13,15-trimethyltetracyclo[8.7.0.0^{2},^{7}.0^{11},^{15}]heptadeca-3,6-dien-5-one
18	(1R,2S,10S,11S,13R,14R,15S,17S)-1-fluoro-14,17-dihydroxy-14-(2-hydroxyacetyl)-2,13,15-trimethyltetracyclo[8.7.0.0^{2},^{7}.0^{11},^{15}]heptadeca-3,6-dien-5-one
19	1-{[4-bromo-2-(2,4-dichlorophenyl)oxolan-2-yl]methyl}-1H-1,2,4-triazole
20	1-[2-(2-chloroethoxy)benzenesulfonyl]-3-(4-methoxy-6-methyl-1,3,5-triazin-2-yl)urea
21	[3-(2-chloro-10H-phenothiazin-10-yl)propyl]dimethylamine
22	1-(2-chlorobenzenesulfonyl)-3-(4-methoxy-6-methyl-1,3,5-triazin-2-yl)urea
23	(3-{14-chloro-2-azatricyclo[9.4.0.0^{3},^{8}]pentadeca-1(11),3,5,7,12,14-hexaen-2-yl}propyl)dimethylamine
24	(1S,2R,10R,11S,14R,15S)-14-hydroxy-14-(2-hydroxyacetyl)-2,15-dimethyltetracyclo[8.7.0.0^{2},^{7}.0^{11},^{15}]heptadec-6-en-5-one
25	(1S,2R,10S,11S,14S,15S,17S)-17-hydroxy-14-(2-hydroxyacetyl)-2,15-dimethyltetracyclo[8.7.0.0^{2},^{7}.0^{11},^{15}]heptadec-6-en-5-one
26	2-[(1S,2R,10S,11S,14R,15S)-14-hydroxy-2,15-dimethyl-5,17-dioxotetracyclo[8.7.0.0^{2},^{7}.0^{11},^{15}]heptadec-6-en-14-yl]-2-oxoethyl acetate
27	2-(4-chlorophenyl)-3-cyclopropyl-1-(1H-1,2,4-triazol-1-yl)butan-2-ol
28	7-hydroxy-3-(4-hydroxyphenyl)-4H-chromen-4-one
29	N,N-diethyl-3-methylbenzamide
30	7,8-dihydroxy-6-methoxy-3-(4-methoxyphenyl)-4H-chromen-4-one
31	5,11-dimethyl-6H-pyrido[4,3-b]carbazole
32	1-{[(2S,3R)-3-(2-chlorophenyl)-2-(4-fluorophenyl)oxiran-2-yl]methyl}-1H-1,2,4-triazole
33	1-{2-[(4-tert-butylphenyl)methyl]propyl}piperidine
34	2-(4-{1-hydroxy-4-[4-(hydroxydiphenylmethyl)piperidin-1-yl]butyl}phenyl)-2-methylpropanoic acid
35	3-(4,6-dimethoxypyrimidin-2-yl)-1-{[3-(trifluoromethyl)pyridin-2-yl]sulfonyl}urea
36	8-[4,4-bis(4-fluorophenyl)butyl]-1-phenyl-1,3,8-triazaspiro[4.5]decan-4-one
37	1-(2-fluorophenyl)-1-(4-fluorophenyl)-2-(1H-1,2,4-triazol-1-yl)ethan-1-ol
38	2-({[(4,6-dimethoxypyrimidin-2-yl)carbamoyl]amino}sulfonyl)-3-formamido-N,N-dimethylbenzamide

39	7-hydroxy-3-(4-methoxyphenyl)-4H-chromen-4-one
40	methyl 2-[N-(2,6-dimethylphenyl)-1-furan-2-ylformamido]propanoate
41	3,5,7-trimethoxy-2-phenyl-4H-chromen-4-one
42	5,7-dihydroxy-3-(4-hydroxyphenyl)-4H-chromen-4-one
43	(1R,2R,5R,8R,9S,10R,12S)-12-hydroxy-11-methyl-6-methylidene-16-oxo-15-oxapentacyclo[9.3.2.1^{5,8}.0^{1,10}.0^{2,8}]heptadecane-9-carboxylic acid
44	(1R,2R,5R,8R,9S,10R,12S)-12-hydroxy-11-methyl-6-methylidene-16-oxo-15-oxapentacyclo[9.3.2.1^{5,8}.0^{1,10}.0^{2,8}]heptadec-13-ene-9-carboxylic acid
45	(1R,2R,5R,8R,9S,10R)-11-methyl-6-methylidene-16-oxo-15-oxapentacyclo[9.3.2.1^{5,8}.0^{1,10}.0^{2,8}]heptadecane-9-carboxylic acid
46	4-[4-(4-chlorophenyl)-4-hydroxypiperidin-1-yl]-1-(4-fluorophenyl)butan-1-one
47	methyl 5-chloro-3-({[(4,6-dimethoxypyrimidin-2-yl)carbamoyl]amino}sulfonyl)-2-methyl-2,3-dihydro-1H-pyrazole-4-carboxylate
48	methyl 3-({[(4-methoxy-6-methyl-1,3,5-triazin-2-yl)carbamoyl]amino}sulfonyl)thiophene-2-carboxylate
49	2-(3,4-dimethoxyphenyl)-3,5,6,7-tetramethoxy-4H-chromen-4-one
50	5,7-dihydroxy-2-(3,4,5-trihydroxyphenyl)-4H-chromen-4-one
51	(1S,2R,10S,11S,14R,15S,17S)-14,17-dihydroxy-14-(2-hydroxyacetyl)-2,15-dimethyltetracyclo[8.7.0.0^{2,7}.0^{11,15}]heptadec-6-en-5-one
52	4-[2-(4-benzylpiperidin-1-yl)-1-hydroxypropyl]phenol
53	1-[2-(2,4-dichlorophenyl)-2-(prop-2-en-1-yloxy)ethyl]-1H-imidazole
54	(3-{2-azatricyclo[9.4.0.0^{3,8}]pentadeca-1(11),3,5,7,12,14-hexaen-2-yl}propyl)dimethylamine
55	1H-indole-3-carboxylic acid
56	2-(1H-indol-3-yl)acetic acid
57	2-(1H-indol-3-yl)acetonitrile
58	1H-indol-3-ylmethanol
59	1H-indole-3-carbaldehyde
60	3-(1H-indol-3-yl)-2-oxopropanoic acid
61	1H-indole-4-carboxylic acid
62	1H-indole-5-carboxylic acid
63	3-(1H-indol-3-yl)propanoic acid
64	5,7-dihydroxy-3-(3-hydroxy-4,5-dimethoxyphenyl)-6-methoxy-4H-chromen-4-one
65	3-(2,3-dihydroxy-4-methoxyphenyl)-7-hydroxy-4H-chromen-4-one
66	{[(2,6-dimethylphenyl)carbamoyl]methyl}triethylazanium
67	2-[(2R,6S)-6-[(2S)-2-hydroxy-2-phenylethyl]-1-methylpiperidin-2-yl]-1-phenylethan-1-one
68	4-[4-(4-chlorophenyl)-4-hydroxypiperidin-1-yl]-N,N-dimethyl-2,2-diphenylbutanamide
69	1-{2-[4-(4-fluorobenzoyl)piperidin-1-yl]ethyl}-3,3-dimethyl-2,3-dihydro-1H-indol-2-one
70	(2R,17R)-12-hydroxy-2,6,6,14,17-pentamethyl-10-oxatetracyclo[7.7.1.0^{2,7}.0^{13,17}]heptadeca-3,13-diene-5,11-dione
71	methyl 2-({[(4,6-dimethoxypyrimidin-2-yl)carbamoyl]amino}sulfonyl)-4-(methanesulfonamidomethyl)benzoate
72	methyl 2-[N-(2,6-dimethylphenyl)-2-methoxyacetamido]propanoate
73	2-chloro-N-(2-ethyl-6-methylphenyl)-N-(1-methoxypropan-2-yl)acetamide
74	methyl 2-({[(4-methoxy-6-methyl-1,3,5-triazin-2-yl)carbamoyl]amino}sulfonyl)benzoate
75	2-(2,4-dihydroxyphenyl)-3,5,7-trihydroxy-4H-chromen-4-one
76	2-(4-chlorophenyl)-2-(1H-1,2,4-triazol-1-ylmethyl)hexanenitrile
77	N-butylbenzamide
78	1-ethyl-2,3-dihydro-1H-indol-2-one
79	N-heptylbenzamide

80	N-hexylbenzamide
81	2-(3,4-dimethoxyphenyl)-5,6,7,8-tetramethoxy-4H-chromen-4-one
82	N-pentylbenzamide
83	N-propylbenzamide
84	oxetan-3-yl 2-({[(4,6-dimethylpyrimidin-2-yl)carbamoyl]amino}sulfonyl)benzoate
85	2,3-dihydro-1H-indol-2-one
86	(2S,3S)-1-(4-chlorophenyl)-4,4-dimethyl-2-(1H-1,2,4-triazol-1-yl)pentan-3-ol
87	5-hydroxy-6-methoxy-2-(4-methoxyphenyl)-7-{[(2S,3R,4S,5S,6R)-3,4,5-trihydroxy-6-({[(2R,3R,4R,5R,6S)-3,4,5-trihydroxy-6-methyloxan-2-yl]oxy}methyl)oxan-2-yl]oxy}-4H-chromen-4-one
88	1-[2-(2,4-dichlorophenyl)pentyl]-1H-1,2,4-triazole
89	2-{4-[3-(2-chloro-10H-phenothiazin-10-yl)propyl]piperazin-1-yl}ethan-1-ol
90	(1S,2R,10S,11S,14R,15S)-14-hydroxy-14-(2-hydroxyacetyl)-2,15-dimethyltetracyclo[8.7.0.0^{2},^{7}.0^{11},^{15}]heptadeca-3,6-diene-5,17-dione
91	5-hydroxy-2-phenyl-4H-chromen-4-one
92	2-chloro-10-[3-(4-methylpiperazin-1-yl)propyl]-10H-phenothiazine
93	4-(dipropylcarbamoyl)-4-(phenylformamido)butanoic acid
94	1-{10-[3-(dimethylamino)propyl]-10H-phenothiazin-2-yl}propan-1-one
95	2-(3,4-dihydroxyphenyl)-3,5,7-trihydroxy-4H-chromen-4-one
96	N-(2,6-dimethylphenyl)-2-{4-[2-hydroxy-3-(2-methoxyphenoxy)propyl]piperazin-1-yl}acetamide
97	4-[(1R,2S)-2-[(4-benzylpiperidin-1-yl)methyl]-1-hydroxypropyl]phenol
98	8-[4-(4-fluorophenyl)-4-oxobutyl]-1-phenyl-1,3,8-triazaspiro[4.5]decan-4-one
99	3,5,6,7-tetramethoxy-2-(4-methoxyphenyl)-4H-chromen-4-one
100	1-(4-benzoylpiperazin-1-yl)propan-1-one
101	1-(4-chlorophenyl)-4,4-dimethyl-3-(1H-1,2,4-triazol-1-ylmethyl)pentan-3-ol
102	1-(4-tert-butylphenyl)-4-[4-(hydroxydiphenylmethyl)piperidin-1-yl]butan-1-ol
103	2-(2,6-dioxopiperidin-3-yl)-2,3-dihydro-1H-isoindole-1,3-dione
104	10-[2-(1-methylpiperidin-2-yl)ethyl]-2-(methylsulfanyl)-10H-phenothiazine
105	(1R,2S,10S,11S,13R,14S,15S,17S)-1-fluoro-13,14,17-trihydroxy-14-(2-hydroxyacetyl)-2,15-dimethyltetracyclo[8.7.0.0^{2},^{7}.0^{11},^{15}]heptadeca-3,6-dien-5-one
106	10-[3-(4-methylpiperazin-1-yl)propyl]-2-(trifluoromethyl)-10H-phenothiazine
107	1-(4-fluorophenyl)-4-{4-hydroxy-4-[3(trifluoromethyl)phenyl]piperidin-1-yl}butan-1-one
108	dimethyl({3-[2-(trifluoromethyl)-10H-phenothiazin-10-yl]propyl})amine
109	1-cyclohexyl-1-phenyl-3-(piperidin-1-yl)propan-1-ol
110	(3-{2-azatricyclo[9.4.0.0^{3},^{8}]pentadeca-1(11),3,5,7,12,14-hexaen-2-yl}-2-methylpropyl)dimethylamine
111	2-(1H-indol-3-yl)ethan-1-ol
112	N,N-diethyl-4-hydroxy-3-methoxybenzamide

Table 2.5. Retention factor (*k*) of compounds in Dataset 1 on ten columns (C1-C10)

ID	Retention factor (*k*)									
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	4.73	3.81	3.05	3.02	0.94	3.24	3.37	3.27	3.39	1.80
2	7.73	6.35	5.04	4.92	1.50	5.42	5.60	5.45	5.65	2.96
3	12.30	10.23	8.13	7.85	2.34	8.85	9.02	8.89	9.06	4.70
4	12.79	10.86	8.39	8.09	2.39	9.20	9.46	9.25	9.53	4.93
5	20.99	17.66	13.90	13.27	3.85	15.42	15.63	15.28	15.70	7.96
6	35.65	30.06	23.60	22.23	6.34	26.55	26.67	26.24	26.85	13.40
7	11.19	9.08	7.24	7.13	2.15	7.76	8.17	7.93	8.18	4.26
8	12.88	10.64	8.32	8.07	2.41	9.06	9.40	9.04	9.02	4.91
9	13.18	10.89	8.47	8.24	2.47	9.23	9.57	9.38	9.68	5.02
10	17.82	13.96	11.40	11.12	3.30	12.39	12.94	12.91	12.91	6.56
11	8.95	7.19	5.71	5.60	1.71	6.12	6.40	6.21	6.44	3.37
12	1.77	1.35	1.16	1.20	0.40	1.20	1.29	1.22	1.27	0.69
13	3.32	2.44	2.15	2.22	0.72	2.20	2.38	2.25	2.34	1.24
14	5.21	3.94	3.45	3.55	1.11	3.56	3.82	3.64	3.77	1.95
15	5.26	3.86	3.45	3.56	1.13	3.58	3.87	3.67	3.80	1.96
16	0.56	0.42	0.42	0.45	0.16	0.39	0.42	0.40	0.42	0.24
17	4.20	3.28	2.74	2.75	0.87	2.86	3.04	2.91	3.03	1.59
18	0.85	0.65	0.58	0.61	0.22	0.58	0.64	0.60	0.63	0.35
19	1.60	1.25	1.12	1.19	0.41	1.11	1.21	1.15	1.20	0.66
20	3.69	2.95	2.72	2.81	0.94	2.65	2.87	2.74	2.84	1.55
21	1.14	0.86	0.74	0.77	0.27	0.76	0.84	0.79	0.83	0.47
22	2.14	1.61	1.37	1.43	0.49	1.43	1.57	1.49	1.56	0.87
23	1.44	1.07	0.91	0.93	0.37	0.91	1.09	1.06	1.08	0.58
24	1.14	0.82	0.72	0.76	0.28	0.74	0.83	0.78	0.81	0.47
25	3.35	2.61	2.25	2.30	0.78	2.34	2.54	2.42	2.52	1.35
26	11.72	9.20	7.71	7.82	2.38	8.05	8.87	8.53	8.67	4.42
27	1.66	1.31	1.14	1.17	0.38	1.15	1.21	1.15	1.20	0.66
28	3.63	2.79	2.44	2.49	0.79	2.48	2.65	2.52	2.62	1.33
29	2.39	1.78	1.59	1.66	0.54	1.60	1.73	1.63	1.70	0.91
30	1.41	1.17	0.92	0.96	0.36	0.92	1.07	1.04	1.08	0.58
31	2.17	1.65	1.50	1.55	0.51	1.48	1.59	1.51	1.57	0.84
32	8.13	6.10	5.48	5.56	1.72	5.65	6.03	5.74	5.93	3.01
33	10.81	8.13	7.35	7.45	2.26	7.66	8.15	7.74	7.98	3.99
34	13.34	9.82	8.91	9.04	2.74	9.29	9.98	9.16	9.77	4.83
35	9.73	6.98	6.56	6.75	2.08	6.76	7.35	6.92	7.13	3.51
36	12.85	9.04	8.65	8.91	2.71	9.00	9.91	9.51	9.51	4.53
37	9.82	7.21	6.67	6.67	2.09	6.90	7.41	7.03	7.24	3.59
38	12.50	8.97	8.34	8.55	2.62	8.67	9.46	8.97	9.14	4.47
39	0.75	0.53	0.61	0.69	0.24	0.52	0.60	0.56	0.58	0.31
40	0.78	0.55	0.62	0.69	0.25	0.55	0.63	0.59	0.61	0.32
41	1.33	0.99	0.90	0.94	0.34	0.91	1.00	0.95	0.99	0.54
42	1.59	1.22	1.10	1.13	0.44	1.10	1.24	1.19	1.23	0.69
43	2.81	2.00	1.95	2.05	0.70	1.92	2.11	1.99	2.05	1.07
44	1.62	1.02	1.13	1.20	0.40	1.09	1.19	1.11	1.16	0.62

45	0.11	0.10	0.18	0.20	0.06	0.09	0.08	0.07	0.08	0.04
46	0.49	0.47	0.77	0.75	0.39	0.68	0.60	0.70	0.72	0.58
47	0.20	0.21	0.39	0.38	0.22	0.34	0.27	0.33	0.35	0.30
48	0.10	0.11	0.23	0.23	0.13	0.20	0.15	0.19	0.20	0.19
49	0.42	0.39	0.65	0.64	0.34	0.58	0.50	0.58	0.61	0.49
50	0.18	0.18	0.35	0.34	0.18	0.31	0.23	0.28	0.30	0.25
51	3.49	3.01	2.81	2.71	0.91	2.66	2.76	2.72	2.80	1.59
52	5.98	5.27	4.84	4.61	1.53	4.61	4.79	4.71	4.88	2.74
53	10.30	9.27	8.38	7.89	2.56	8.02	8.32	8.20	8.51	4.72
54	1.15	0.91	0.86	0.79	0.26	0.82	0.86	0.83	0.86	0.49
55	2.49	1.88	1.77	1.77	0.58	1.70	1.81	1.74	1.79	0.96
56	7.66	5.82	5.35	5.20	1.85	5.18	5.81	5.57	5.74	3.05
57	12.71	10.21	9.02	8.63	2.99	8.93	9.89	9.57	9.84	5.25
58	3.02	2.28	2.12	2.12	0.78	2.07	2.34	2.23	2.30	1.25
59	4.95	3.14	2.96	3.12	1.32	2.23	2.76	2.55	2.59	1.48
60	5.89	5.08	4.42	4.38	1.58	4.29	4.70	4.56	4.75	2.71
61	9.71	8.63	7.36	7.21	2.57	7.18	7.82	7.66	7.96	4.53
62	16.07	14.76	12.39	11.97	4.23	12.11	13.15	13.30	13.49	7.64
63	0.72	0.53	0.51	0.54	0.22	0.46	0.53	0.50	0.52	0.30
64	0.54	0.37	0.34	0.36	0.16	0.29	0.36	0.33	0.34	0.20
65	0.99	0.73	0.71	0.75	0.31	0.62	0.73	0.68	0.70	0.40
66	1.40	1.08	0.96	0.96	0.37	0.94	1.05	1.00	1.04	0.59
67	0.83	0.62	0.57	0.59	0.24	0.58	0.62	0.58	0.61	0.35
68	2.88	2.07	1.87	1.95	0.65	1.91	2.10	1.97		1.08
69	3.01	2.20	1.97	2.05	0.68	2.02	2.22	2.10		1.14
70	4.92	3.82	3.33	3.39	1.07	3.47	3.67	3.49		1.89
71	17.66	14.42	11.27	11.14	3.26	12.50	13.15	12.74		6.67
72	9.73	7.15	6.38	6.53	2.02	6.71	7.19	6.81		3.55
73	14.79	11.14	9.59	9.75	2.94	10.23	11.02	10.50		5.42
74	3.69	2.81	2.51	2.58	0.86	2.58	2.82	2.69		1.47
75	2.48	1.88	1.63	1.69	0.59	1.69	1.88	1.79		1.00
76	1.02	0.69	0.67	0.71	0.27	0.66	0.75	0.70		0.40
77	7.41	5.52	5.04	5.15	1.62	5.22	5.57	5.26		2.78
78	7.96	5.60	5.24	5.46	1.72	5.41	5.86	5.48		2.84
79	3.40	2.60	2.38	2.44	0.79	2.42	2.58	2.45		1.33
80	13.09	12.22	9.82	9.66	2.79	10.86	11.17	10.86		5.64
81	0.46	0.36	0.51	0.54	0.17	0.34	0.35	0.33		0.21
82	1.53	1.12	0.99	1.04	0.37	1.01	1.13	1.06		0.61
83	1.36	1.00	0.88	0.93	0.34	0.89	1.00	0.94		0.54
84	1.90	1.71	1.21	1.26	0.46	1.17	1.34	1.26		0.70
85	2.36	1.81	1.55	1.60	0.55	1.57	1.74	1.64		0.90
86	2.51	2.20	0.62	0.73	0.31	0.64	1.21	1.14		0.48
87	0.46	0.32	0.31	0.32	0.14	0.30	0.35	0.32		0.19
88	4.78	3.30	3.08	3.23	1.08	3.14	3.60	3.37		1.74
89	16.33	12.76	10.91	10.94	3.33	11.86	12.53	11.97		6.18
90	5.83	4.61	3.82	3.79	1.16	4.18	4.36	4.26		2.20

Table 2.6. Retention factor (*k*) of compounds in Dataset 2 on five columns (C11-C15)

ID	Retention factor (*k*)				
	C11	C12	C13	C14	C15
1	0.50	0.44	0.41	0.38	0.45
2	1.55	1.37	1.21	1.06	1.22
3	4.67	4.15	3.52	2.76	3.21
4	0.22	0.20	0.19	0.18	0.21
5	0.75	0.64	0.58	0.53	0.61
6	1.07	0.94	0.81	0.72	0.84
7	3.27	3.13	2.39	1.86	2.41
8	5.61	5.04	3.97	3.02	3.83
9	9.71	8.89	6.78	4.76	5.70
10	0.17	0.13	0.10	0.11	0.11
11	0.41	0.31	0.28	0.28	0.33
12	3.05	2.32	2.12	1.82	2.14
13	0.20	0.15	0.10	0.12	0.14
14	0.46	0.35	0.29	0.31	0.34
15	1.27	1.11	1.05	0.92	1.09
16	2.70	2.38	2.14	1.73	2.06
17	4.58	4.06	3.58	2.73	3.34
18	7.93	7.16	4.90	4.36	5.38
19	1.57	1.38	1.29	1.08	1.30
20	2.70	2.41	2.20	1.75	2.10
21	1.11	1.04	0.88	0.72	0.86
22	4.41	4.34	3.53	2.42	3.04
23	14.59	14.55	11.40	6.67	8.61
24	0.30	0.26	0.22	0.20	0.24
25	0.33	0.29	0.28	0.26	0.31
26	0.61	0.53	0.51	0.46	0.55
27	1.85	1.60	1.50	1.23	1.49
28	3.17	2.78	2.51	1.98	2.43
29	9.55	8.63	7.24	5.01	6.28
30	0.66	0.55	0.52	0.46	0.56
31	0.55	0.46	0.48	0.45	0.52
32	1.56	1.32	1.30	1.13	1.36
33	2.64	2.25	2.18	1.80	2.15
34	4.40	3.78	3.59	2.84	3.44
35	7.45	6.50	6.03	4.48	5.45
36	12.76	11.35	10.23	7.05	8.65
37	1.33	1.13	1.14	1.00	1.21
38	2.24	1.92	1.88	1.59	1.91
39	3.77	3.26	3.12	2.51	3.03
40	1.59	1.45	1.36	1.15	1.41
41	2.81	2.59	2.34	1.89	2.33
42	2.15	1.96	1.80	1.46	1.78
43	0.62	0.52	0.55	0.49	0.58
44	0.81	0.68	0.70	0.62	0.73

45	1.18	0.99	0.99	0.86	1.00
46	1.56	1.32	1.27	1.06	1.28
47	0.44	0.35	0.39	0.36	0.42
48	2.72	2.40	2.21	1.71	2.08
49	4.43	3.91	3.63	2.62	3.31
50	3.09	2.82	2.52	1.94	2.35
51	1.65	1.39	1.34	1.09	1.32
52	3.01	2.60	2.44	1.84	2.24
53	6.07	5.11	4.90	3.40	4.46
54	1.80	1.50	1.49	1.26	1.53
55	2.90	2.43	2.36	1.86	2.28
56	1.79	1.47	1.51	1.28	1.54
57	2.45	2.07	2.02	1.63	2.00
58	4.06	3.47	3.36	2.48	3.16
59	3.59	3.05	2.99	2.27	2.88
60	3.85	3.30	3.21	2.37	3.00
61	4.45	3.66	3.64	2.73	3.44
62	3.91	3.23	3.20	2.42	3.08
63	3.37	3.15	2.85	2.17	2.70
64	5.48	5.35	4.54	3.18	4.03
65	6.34	6.21	5.22	3.55	4.51
66	8.26	8.15	6.70	4.36	5.58
67	5.50	4.89	4.50	3.30	4.12
68	9.29	9.38	7.59	4.89	6.31
69	10.81	10.96	8.75	5.47	7.05
70	9.51	9.62	7.73	4.91	6.37
71	3.37	3.19	2.84	2.08	2.58
72	5.57	5.45	4.61	3.16	4.02
73	9.06	8.87	7.36	4.80	6.11
74	15.70	15.67	12.50	7.59	9.84
75	27.10	27.29	21.18	11.94	15.96
76	19.36	18.79	15.21	9.23	12.11
77	9.35	9.53	7.62	4.81	6.14
78	15.74	16.07	12.62	7.35	9.59
79	13.58	13.21	10.86	6.68	8.65
80	8.15	8.02	6.59	4.34	5.50
81	21.43	23.93	17.06	9.23	11.97
82	0.82	0.70	0.74	0.67	0.83
83	1.22	1.04	1.07	0.94	1.12
84	1.21	1.03	1.06	0.93	1.12
85	1.38	1.18	1.21	1.04	1.28
86	1.92	1.65	1.66	1.40	1.63
87	1.52	1.31	1.32	1.16	1.43
87	1.52	1.31	1.32	1.16	1.43

Table 2.7. Retention factor (*k*) of compounds in Dataset 3 on Zorbax Eclipse Plus C18 column (C16)

ID	*k*	ID	*k*	ID	*k*	ID	*k*
1	22.91	29	11.48	57	6.46	85	1.12
2	5.62	30	4.57	58	8.32	86	33.88
3	16.60	31	1.38	59	2.00	87	2.14
4	8.13	32	77.62	60	4.90	88	95.50
5	17.78	33	13.49	61	1.58	89	4.79
6	7.59	34	6.61	62	1.48	90	3.02
7	53.70	35	5.50	63	3.98	91	93.33
8	7.59	36	29.51	64	8.91	92	6.76
9	5.62	37	12.30	65	2.63	93	7.94
10	16.22	38	5.37	66	10.47	94	6.76
11	48.98	39	12.02	67	2.95	95	3.16
12	50.12	40	39.81	68	21.38	96	1.32
13	10.96	41	35.48	69	4.37	97	1.70
14	3.31	42	5.89	70	6.31	98	2.19
15	6.17	43	13.18	71	16.98	99	50.12
16	13.80	44	12.02	72	14.13	100	1.26
17	8.51	45	42.66	73	85.11	101	91.20
18	6.61	46	3.39	74	7.59	102	64.57
19	47.86	47	64.57	75	2.24	103	1.66
20	9.55	48	6.17	76	66.07	104	19.05
21	9.77	49	30.20	77	6.76	105	1.26
22	10.00	50	3.31	78	4.27	106	12.88
23	11.75	51	3.24	79	75.86	107	5.62
24	9.12	52	2.09	80	33.88	108	17.38
25	8.13	53	5.50	81	28.18	109	5.25
26	18.20	54	5.25	82	15.14	110	7.94
27	42.66	55	1.58	83	3.09	111	2.24
28	2.40	56	8.71	84	5.50	112	1.86

2.1.2 Database for non-targeted metabolomics

In this thesis, the application of retention prediction in Non-Targeted Metabolomics (NTM) is also evaluated. To achieve this, a database of compounds with information about molecular weight and retention was required. A database originally used by Hall *et al*. to derive QSRR models by artificial neural networks (ANN), consisting of 1882 compounds for which retention index information and mass values are reported was utilised [4]. In the original data collection, compound mixtures were analysed on an HPLC/MS system, where an Agilent 1100 HPLC (Agilent, Santa Clara, CA, USA) system interfaced to a QTOF-2 mass spectrometer (Waters Associates, Beverly, MA, USA). A Zorbax SBC18 (1 mm × 150 mm, 3.5 um particle size) column containing a Zorbax Stable Bond (1 mm × 17 mm, 5 mm particle size) OptiGuard precolumn was used for the separation [4]. Solvent A was 0.01% heptafluorobutryic acid

(HFBA) in water and solvent B was 0.01% HFBA, in 90:10 acetonitrile:water. The flow-rate was 75 μL/min with a linear solvent gradient from 0% solvent B to 100% solvent B in 17 min, followed by a 5-min isocratic hold at 100% B. Finally, retention index (RI) values of compounds in the database were determined experimentally with a wide range from 204 to 1041 [4].

The Tanimoto similarity analysis of this database indicates a large diversity in the 1882 compounds although some isomers were found. To evaluate the performance of the established QSRR models in eliminating false positives, groups of representative compounds with the same mass values were selected. In all, 34 groups of compounds (248 compounds in total) were chosen with at least five compounds having the same mass value in each group. The 1882 compounds can be found in Appendix 1, and the 248 representative compounds are listed in Table 2.8.

2.2 Data collection

2.2.1 Sample preparation

Most of the retention information for compounds used for RPLC QSRR modelling in this study was obtained from literature [1, 4]. However, to evaluate the predictive ability of the proposed QSRR approach, five new compounds including an acid, a base and three neutrals that never been used in the modelling process were selected and tested on five new reversed-phase columns. Therefore, five analytical grade compounds were purchased from Sigma-Aldrich (St. Louis, MO, USA): pindolol, 8-dihydroxynaphthalene, 4-ethylnitrobeneze, 2-phenylbutane and 4-heptylbenzoic acid. HPLC grade Acetonitrile and methanol were supplied by VWR International (Melbourne, VIC, Australia) and Sigma-Aldrich (St. Louis, MO, USA), respectively. 18.2 MΩ Milli-Q water produced using a Millipore Gradient water purification (Millipore, Bedford, MA, USA) system, was used to prepare mobile phase and sample solutions. Standard stock solutions (1000 μg/mL) of each compound were obtained by dissolving an appropriate amount of each standard in acetonitrile-water (50:50) solution.

2.2.2 Instrumentation

All experiments were performed using a Thermo Fisher Scientific Ultimate 3000 instrument (Lane Cove, Australia) equipped with a DGP-3600RS pump, a DAD-3000RS diode array detector, a WPS-3000TRS autosampler with temperature control, and a TCC-3000RS column compartment. Chromeleon software (ver. 7.1.2) was used for system control and data processing.

Table 2.8. 34 groups (248) of representative compounds used for the elimination of false positives

ID	Name	MW	RI
111	N,N-dimethylallylamine	85.1486	459
142	(R)-(-)-2-methylpyrrolidine	85.1486	474
213	cyclopentylamine	85.1486	509
1493	1-methylpyrrolidine	85.1486	428
1508	piperidine	85.1486	492
137	N,N-dimethylisopropylamine	87.1644	472
306	1,2-dimethylpropylamine	87.1644	538
335	1-ethylpropylamine	87.1644	542
352	2-aminopentane	87.1644	545
1513	N-methylbutylamine	87.1644	534
67	4-aminobutan-2-ol	89.137	381
75	2-amino-2-methylpropan-1-ol	89.137	401
79	(2-methoxyethyl)(methyl)amine	89.137	408
113	3-methoxypropan-1-amine	89.137	461
1597	2-dimethylaminoethanol	89.137	343
101	3-aminopyridine	94.1158	447
108	4-aminopyridine	94.1158	456
132	3-methylpyridazine	94.1158	470
157	2-aminopyridine	94.1158	484
288	4-methylpyrimidine	94.1158	533
1561	1-vinylimidazole	94.1158	474
181	triethylamine	101.191	497
333	N,N-dimethylbutylamine	101.191	541
340	N-ethylbutylamine	101.191	543
544	1,3-dimethylbutylamine	101.191	589
1512	diisopropylamine	101.191	524
48	2-methoxypropanamide	103.121	334
55	gamma aminobutyric acid	103.121	344
1608	dimethylglycine	103.121	246
1636	L-alpha-aminobutyric_acid	103.121	278
1638	2-aminoisobutyric_acid	103.121	287
1703	glycine_ethyl_ester	103.121	475
200	2,6-lutidine	107.155	506
237	2,5-lutidine	107.155	517
244	2,3-lutidine	107.155	518
248	2-ethylpyridine	107.155	519
262	2,4-lutidine	107.155	527
271	3,4-lutidine	107.155	528
282	3,5-lutidine	107.155	532
296	3-ethylpyridine	107.155	535
299	4-ethylpyridine	107.155	536
404	N-methylaniline	107.155	557
501	o-toluidine	107.155	577
511	p-toluidine	107.155	579

1514	benzylamine	107.155	552
1563	m-toluidine	107.155	577
77	p-phenylenediamine	108.143	404
122	m-phenylenediamine	108.143	465
206	4-(aminomethyl)pyridine	108.143	507
218	2-(methylamino)pyridine	108.143	511
219	4-(methylamino)pyridine	108.143	511
247	o-phenylenediamine	108.143	518
443	2,6-dimethylpyrazine	108.143	568
458	2,5-dimethylpyrazine	108.143	570
468	4,6-dimethylpyrimidine	108.143	572
536	2-ethylpyrazine	108.143	587
211	1-ethylpiperidine	113.202	509
425	heptamethyleneimine	113.202	563
449	N-methylcyclohexylamine	113.202	568
554	cycloheptylamine	113.202	593
614	cyclohexanemethylamine	113.202	609
162	1,4-dimethylpiperazine	114.190	486
205	4-(aminomethyl)piperidine	114.190	507
221	1-methylhomopiperazine	114.190	512
229	(R)-(+)-3-(dimethylamino)pyrrolidine	114.190	515
236	trans-2,5-dimethylpiperazine	114.190	516
243	1-ethylpiperazine	114.190	518
250	2,6-dimethylpiperazine	114.190	520
279	2-(aminomethyl)piperidine	114.190	531
322	1,2-diaminocyclohexane	114.190	540
442	N-benzylmethylamine	121.182	567
491	N,N-dimethylaniline	121.182	576
522	N-ethylaniline	121.182	583
552	2-methylbenzylamine	121.182	592
567	3-methylbenzylamine	121.182	596
568	4-methylbenzylamine	121.182	596
604	2,3-dimethylaniline	121.182	607
636	N-methyl-o-toluidine	121.182	615
658	3,5-dimethylaniline	121.182	622
1504	2,4,6-collidine	121.182	544
1517	1-phenylethylamine	121.182	578
1520	phenethylamine	121.182	595
212	2,6-diaminotoluene	122.169	509
270	2-(dimethylamino)pyridine	122.169	528
273	2-(2-pyridyl)ethylamine	122.169	529
277	2,4-diaminotoluene	122.169	530
415	2,3-diaminotoluene	122.169	560
435	3,4-diaminotoluene	122.169	566
1562	4-dimethylaminopyridine	122.169	533
367	N-isopropyl-N-methyl-tert-butylamine	129.245	548
535	N,N-diethylbutylamine	129.245	586
600	diisobutylamine	129.245	606

687	dibutylamine	129.245	628
714	N,N-dimethylhexylamine	129.245	638
850	2-amino-6-methylheptane	129.245	691
1496	N,N-diisopropylethylamine	129.245	546
1518	di-sec-butylamine	129.245	586
1526	octylamine	129.245	727
115	(1,4-dioxan-2-ylmethyl)methylamine	131.174	462
201	(3S,4R)-4-ethoxyoxolan-3-amine_hydrochloride	131.174	507
1649	L-isoleucine	131.174	482
1651	leucine	131.174	494
1652	L-norleucine	131.174	502
1653	6-aminocaproic_acid	131.174	506
1655	beta-leucine	131.174	521
323	5-aminoindole	132.165	540
395	1-methylbenzimidazole	132.165	556
420	2-methylbenzimidazole	132.165	562
486	6-aminoindole	132.165	575
537	5-methylbenzimidazole	132.165	587
613	7-aminoindole	132.165	609
483	1,2,3,4-tetrahydroquinoline	133.193	575
520	2-methylindoline	133.193	582
530	1,2,3,4-tetrahydroisoquinoline	133.193	586
550	1-aminoindan	133.193	592
623	4-aminoindan	133.193	612
652	5-aminoindan	133.193	621
882	N-allylaniline	133.193	706
473	N,N-dimethylbenzylamine	135.208	573
488	N,N-dimethyl-o-toluidine	135.208	575
523	N-ethylbenzylamine	135.208	583
571	(S)-(-)-N,alpha-dimethylbenzylamine	135.208	598
584	N-methyl-phenethylamine	135.208	601
597	3-methyl-N-methylbenzylamine	135.208	605
601	N-isopropylaniline	135.208	606
618	N,N-dimethyl-p-toluidine	135.208	610
649	N,N-dimethyl-m-toluidine	135.208	620
676	N-ethyl-m-toluidine	135.208	626
690	N-ethyl-o-toluidine	135.208	629
804	4-isopropylaniline	135.208	671
806	4-propylaniline	135.208	674
821	2,4,6-trimethylaniline	135.208	681
845	o-isopropylaniline	135.208	689
879	2-propylaniline	135.208	705
1521	L-amphetamine	135.208	612
254	N-ethyl-4-pyridinemethylamine	136.196	521
263	3-(2-pyridyl)propylamine	136.196	527
285	N,N-dimethyl-p-phenylenediamine	136.196	532
295	2-(2-methylaminoethyl)pyridine	136.196	535
331	m-xylylenediamine	136.196	541

334	o-xylylenediamine	136.196	541
533	2,3,5,6-tetramethylpyrazine	136.196	586
587	4,5-dimethyl-1,2-phenylenediamine	136.196	602
89	3-amino-2-methylbenzamide	150.180	426
147	(4-aminophenyl)-N-methylcarboxamide	150.180	478
220	N-(2-aminophenyl)acetamide	150.180	511
233	2-amino-2-phenylacetamide	150.180	516
444	2-amino-N-phenylacetamide	150.180	568
112	2-(4-hydroxyphenyl)acetamide	151.165	461
126	2-hydroxy-2-phenylacetamide	151.165	467
308	2-(aminomethyl)benzoic_acid	151.165	538
424	5-amino-2-methylbenzoic_acid	151.165	563
434	2,5-dimethylpyrrole-3,4-dicarbaldehyde	151.165	566
775	(2-hydroxyphenyl)-N-methylcarboxamide	151.165	660
1648	2-phenylglycine	151.165	458
1679	acetaminophen	151.165	494
1689	N-phenylglycine	151.165	599
496	2-(ethylamino)-4-methylphenol	151.208	577
598	methyl(2-phenoxyethyl)amine	151.208	605
630	3-phenoxypropylamine_chloride	151.208	613
640	4-methoxy-2,3-dimethylphenylamine	151.208	617
737	2-amino-4-propylphenol	151.208	646
951	1-ethyl-2,5-dimethylpyrrole-3-carbaldehyde	151.208	735
1664	N-benzylethanolamine	151.208	560
417	((3S)-3-1,2,3,4-tetrahydroisoquinolyl)methan-1-ol	163.219	562
609	8-methoxy-1,2,3,4-tetrahydroquinoline	163.219	608
785	2,3-dimethyl-5,6,7-trihydroindol-4-one	163.219	665
1006	4-phenylmorpholine	163.219	760
1601	methcathinone	163.219	601
104	3-amino-3-phenylpropanoic_acid	165.191	453
154	2-propylpyridine-4-carboxylic_acid	165.191	482
251	4-(4-pyridyl)butanoic_acid	165.191	520
347	methyl_3-(3-pyridyl)propanoate	165.191	544
432	N-(4-hydroxyphenyl)-N-methylacetamide	165.191	565
498	methyl_2-amino-2-phenylacetate	165.191	577
1605	2-dimethylaminobenzoic_acid	165.191	498
1606	3-dimethylaminobenzoic_acid	165.191	618
1607	4-dimethylaminobenzoic_acid	165.191	731
1656	phenylalanine	165.191	521
1691	3,5-dimethylanthranilic_acid	165.191	745
1692	N-ethylanthranilic_acid	165.191	776
1718	benzocaine	165.191	767
773	4-hydroxy-1-methylhydroquinolin-2-one	175.187	659
932	1-methylindole-3-carboxylic_acid	175.187	727
945	1,7-dimethylbenzo[d]azolidine-2,3-dione	175.187	732
1085	methyl_indole-3-carboxylate	175.187	805
1152	1-methylindole-2-carboxylic_acid	175.187	841
1697	indole-3-acetic_acid	175.187	678

1726	indoxyl_acetate	175.187	834
1834	citrulline	175.187	253
210	1,2,3,4-tetrahydroisoquinoline-3-carboxylic_acid	177.202	509
437	1-(4-hydroxyphenyl)pyrrolidin-2-one	177.202	567
590	N-(3-acetylphenyl)acetamide	177.202	603
1113	4,6-dimethoxyindole	177.202	821
1796	5-hydroxytryptophol	177.202	530
293	N-[(4-hydroxyphenyl)ethyl]acetamide	179.218	534
341	2-amino-3-phenylbutanoic_acid_chloride	179.218	544
477	3-(1,3-dioxan-2-yl)phenylamine	179.218	574
783	6-propyl-2H-benzo[d]1,3-dioxolene-5-ylamine	179.218	663
889	1-cyclohexylazoline-2,5-dione	179.218	709
1251	ethyl_2-(phenylamino)acetate	179.218	891
1693	2-ethylanilinoacetic_acid	179.218	794
831	2-(2-naphthyloxy)ethylamine	187.241	686
832	3-(phenylamino)cyclohex-2-en-1-one	187.241	687
849	2,7,8-trimethylquinolin-4-ol	187.241	691
878	8-ethyl-2-methylquinolin-4-ol	187.241	704
1009	1,2,3-trimethylcyclohepta[1,2-c]pyrrol-6-one	187.241	761
1147	1,2,3,4,9-pentahydro-4aH-carbazol-1-ol	187.241	840
1190	1-phenyl-2-pyrrolylethan-1-ol	187.241	859
622	3,3-dimethyl-2-phenyl-1-pyrrolin-5-ol	189.257	612
1051	(2Z)-3-(dimethylamino)-1-(3-methylphenyl)prop-2-en-1-one	189.257	788
1191	2-(1-ethylindol-3-yl)ethan-1-ol	189.257	859
1396	1,2-dimethyl-1,2,3,4-tetrahydroquinoline-6-carbaldehyde	189.257	977
1436	4-piperidylbenzaldehyde	189.257	1002
1181	8-ethylthioquinoline	189.275	856
192	2-methyl-1,2,3,4-tetrahydroisoquinoline-3-carboxylic_acid	191.229	502
572	3-(3-hydroxypropyl)indolin-2-one	191.229	598
669	3-oxo-N-benzylbutanamide	191.229	625
704	2-acetyl-3-methyl-4-oxo-5,6,7-trihydroindole	191.229	634
787	2-hydroxyphenyl_pyrrolidinyl_ketone	191.229	665
874	1-(4-methoxyphenyl)pyrrolidin-2-one	191.229	701
703	N-(4-methoxy-2,6-dimethylphenyl)acetamide	193.245	633
732	3,4-diacetyl-1,2,5-trimethylpyrrole	193.245	645
745	methyl_3-(3-amino-4-methylphenyl)propanoate	193.245	649
970	N-[2-(4-methylphenoxy)ethyl]acetamide	193.245	747
1354	butyl_4-aminobenzoate	193.245	952
1704	L-phenylalanine_ethyl_ester	193.245	641
605	4-methylpyridino[3,2-h]quinoline	194.235	607
659	acridine-4-ylamine	194.235	622
665	acridine-9-ylamine	194.235	624
731	2-phenylbenzimidazole	194.235	645
927	2-(3-pyridyl)indole	194.235	725
805	5,6,7,8,10-pentahydroacridin-9-one	199.252	671
893	2-methyl-1H,3H-naphtho[1,2-e]1,3-oxazine	199.252	710
896	4-(phenylmethoxy)phenylamine	199.252	711
1130	2-methoxy-5-phenylphenylamine	199.252	831

1350	2,5-dimethyl-1-phenylpyrrole-3-carbaldehyde	199.252	948
1176	5,6,7,8,9-pentahydro-4aH-carbazole-3-carboxylic_acid	215.251	854
1220	3-(5-phenylpyrrol-2-yl)propanoic_acid	215.251	870
1241	2-(2,5-dimethylpyrrolyl)benzoic_acid	215.251	884
1260	1,2,3,4,9-pentahydro-4aH-carbazolecarboxylic_acid	215.251	899
1329	3-(2,5-dimethylpyrrolyl)benzoic_acid	215.251	936
1351	4-(2,5-dimethylpyrrolyl)benzoic_acid	215.251	948
441	N-[4-(4-pyridylmethyl)phenyl]acetamide	226.277	567
722	4-{[4-(dimethylamino)phenyl]azamethylene}cyclohexa-2,5-dien-1-one	226.277	641
1053	(4-aminophenyl)-N-(4-methylphenyl)carboxamide	226.277	789
1129	(3,4-dimethylphenyl)-N-(2-pyridyl)carboxamide	226.277	831
1317	(2-aminophenyl)-N-(4-methylphenyl)carboxamide	226.277	928

The five HPLC columns employed in this study were obtained from Thermo Fisher Scientific: an Acclaim™ 120 C18 column (4.6 × 150 mm, 5.0 μm); an Accucore C18 column (4.6 × 150 mm, 2.6 μm); a Hypersil GOLD C8 column (4.6 × 150 mm, 5.0 μm); a Hypersil GOLD C18 column (4.6 × 150 mm, 5.0 μm) and a Hypersil ODS C18 column (4.6 × 150 mm, 5.0 μm). The mobile phase was prepared exactly as for Wilson's study [1]: For neutral compounds, the mobile phase was acetonitrile-water, mixed on-line at 50% (v/v). For acidic or basic compounds, the mobile phase was acetonitrile-buffer, where the buffer is 31.2 mM potassium phosphate (pH 2.80) prepared by titrating phosphoric acid with KOH; *i.e.*, pH measurements were carried out on the buffer, prior to the addition of acetonitrile. All data were collected using UV detection at 205 nm, with a column temperature of 35 °C, and a flow rate of 1.5 ml/min. Columns were equilibrated with 15-20 column volumes of eluent to guarantee stable equilibrium situations. The void time was measured for each column by the injection of uracil.

2.2.3 Retention data collection

Retention data of the five new compounds and the five new columns are listed in Table 2.9.

Table 2.9. Measured retention time (min) of five new representative compounds on five columns

Compound	t_R	Measured retention times (min)				
		Acclaim™ 120	Accucore	Hypersil GOLD	Hypersil GOLD	Hypersil ODS
uracil (marker)	t_0	0.924	0.824	1.182	1.157	1.015
ethylbezene (reference)	t_{ref}	11.249	6.207	5.315	6.307	6.990
pindolol (C1)	t_1	1.040	0.890	1.290	1.257	1.465
1, 8-dihydroxynaphthalene (C2)	t_2	3.024	1.782	2.390	2.465	2.357
4-ethylnitrobeneze (C3)	t_3	7.882	4.190	4.282	4.882	5.315
2-phenylbutane (C4)	t_4	26.740	14.382	10.049	13.107	15.765
4-heptylbenzoic acid (C5)	t_5	27.249	14.382	10.924	15.299	18.440

2.3 QSRR model generation

2.3.1 Software

MarvinSketch version 16.2.15 from ChemAxon (Budapest, Hungary) was used for drawing the molecular structures [5]. Initial conformational searches to find the 50 lowest energy structures were performed using a Merck Molecular Force Field (MMff94) implemented in Balloon [6-11]. Geometry optimisation using the semi-empirical Parametric Method number 7 (PM7) [12] was performed in Molecular Orbital PACkage (MOPAC) [12], followed by further geometry optimisation of the lowest energy structure using density functional theory implemented in Gaussian 09 [13-15]. Dragon 6.0 (Talete, Milano, Italy) and VolSurf+ 1.0.7.1 (Molecular Discovery Ltd., Hertfordshire, UK) was employed for calculation of molecular descriptors [16]. A genetic algorithm (GA) in Matlab R2013b (The Mathworks Inc., Natick, MA, USA) was utilised to select the most important descriptors and to build the QSRR models for each stationary phase material [17]. Statistical evaluation of the data and multivariate data analysis has also been performed in Matlab. For more information go to the user manual in Section 2.4.

2.3.2 Calculation of molecular descriptors

The Dragon 6.0 (Talete, Milano, Italy) and VolSurf+ 1.0.7.1 (Molecular Discovery Ltd., Hertfordshire, UK) software tools were employed to generate molecular descriptors. The procedure for the generation of molecular descriptors using Dragon was as follows. The structures of the molecules were sketched in MarvinSketch. Initial conformational searches to find the 50 lowest energy structures were performed using Balloon with a Merck Molecular Force Field (MMff94) [6-8]. The lowest energy conformer was taken as the input structure for geometry optimisation using a semi-empirical PM7 method implemented in MOPAC [9, 18], the resulting geometry was further refined with the Gaussian program applying the Becke 3-parameter (exchange) with correlation by Lee Yang and Parr, (B3LYP) [13, 19, 20] functional and the 6-31G-(d) basis set [21]. Optimisations were performed in acetonitrile using the integral equation formalism variant of the polarizable continuum model (IEFPCM) [22]. Following each geometry optimisation, harmonic frequency analysis was carried out to confirm the nature of each stationary point as an equilibrium structure. The resulting minimum energy conformations of the compounds in this study were input into Dragon to calculate molecular descriptors. Dragon software was able to calculate over 4000 molecular descriptors, consisting of constitutional, topological, geometrical, electrostatic, physical, shape, and quantum chemical descriptors. To minimize subsequent problems of chance correlation, descriptors with constant or near constant values, descriptors with a standard deviation less than 0.0001, descriptors which were strongly correlated to other descriptors (using a

correlation coefficient >0.95) and those descriptors not available for all compounds were excluded [23, 24]. After this reduction step, 128 molecular descriptors were obtained. Before statistical analysis, all the descriptors were scaled to zero mean and unit variance (auto-scaling procedure) because the numerical values of the descriptors varied significantly. The resulting descriptor sets were used to build predictive models for the experimental chromatographic retention data.

In terms of the calculation of molecular descriptors using VolSurf+, Molecular descriptors were calculated from the canonical simplified molecular-input line entry system strings (SMILES) of compounds using VolSurf+ at a user-defined pH (pH=2.5) [25, 26]. The VolSurf+ software was able to calculate 128 molecular descriptors based on 3D Molecular Interaction Fields (MIFs) produced by GRID [27, 28], which is a computational tool for determining energetically favourable binding sites on molecules of known structure [27-30]. Vol-Surf+ descriptors are produced from 3D interaction energy grid maps, which are particularly ADME relevant (absorption, distribution, metabolism, and excretion), and are easy to interpret as the information present in 3D maps is compressed into 2D numerical descriptors [28, 29]. In all, 128 molecular descriptors were calculated following the 3D structure conversion and conformational analysis of structures. All descriptors were auto-scaled (*i.e.*, to have zero mean and unit variance) before use for the modelling process. For more information go to the user manual in Section 2.4

2.3.3 Genetic algorithm

In QSRR, retention is expressed as a function of molecular descriptors. Given the huge number of molecular descriptors that can be generated using either Dragon or VolSurf+ software, an appropriate variable selection method is highly desired to choose the most informative descriptors to build models. The GA, introduced by Holland [31], is a stochastic search procedure inspired by the rules of natural selection to select features without making any assumptions about the search space. By relying on bio-inspired operators such as mutation, crossover, and selection, it has been used to generate high-quality solutions to optimisation and search problems [11, 32, 33]. In a GA, each variable is called a gene, and a set of variables is called a chromosome. The relationship is more like a bit and a bit string in genetic terms. Generating an initial population of chromosomes by random choice of variables is the first step of a GA optimisation [25, 34]. After that, pairs of chromosomes are chosen randomly as parents and crossover operations performed so that a new generation of child chromosomes are produced with better fitness. For the last step, a mutation is performed to maintain genetic diversity from the initial random population to the next generations [11, 32]. The cycle of the

evaluation, selection, crossover, and mutation processes is then repeated until a stopping criterion is satisfied.

The GA optimisation relies on a randomly generated initial population, which can potentially limit its capability to find the most relevant variables within a large search domain [34]. Therefore, the final results of replicate runs could be very different. To pick the most informative subset of molecular descriptors to build QSRR models within a reasonable computational time, 100 runs with different initial populations were generated and the frequency with which each variable was selected as the top chromosome of each run was calculated [24, 25, 35]. Moreover, a popular version of a GA-PLS algorithm, originally written by Leardi in Matlab (The Mathworks Inc., Natick, MA, USA) was used for descriptor selection [32, 33]. The parameters of the GA in the present study were: 50 chromosomes in the original population; a maximum of 20 variables per chromosome (average probability of selection of 10 variables), 1% probability of mutation, 50% probability of cross-over, and a backward elimination phase after every 100 evaluations. The minimum number of compounds in the training set was set at five. To cope with the variability of the results arising from the intrinsic random selection nature of the GA, the GA-PLS modelling was repeated five times and the results were averaged [24, 26].

2.3.4 Partial least square

An excessive number of molecular descriptors has been generated to describe the molecules in the retention databases used in this thesis, and many of those generated descriptors are redundant, co-linear and can be regarded as noise [33, 36, 37]. In this case, the use of an appropriate descriptor selection method like multiple linear regression (MLR) or partial least squares regression (PLS) is highly desirable to reduce the risk of over-fitting and chance correlation by excluding the noise from the model [24-26, 34, 35]. As a linear multiple regression method, PLS has been used commonly in in chemometric and multivariate calibration solutions [33, 36].

PLS can deal with databases which contain more variables than the number of samples. Also, it is particularly helpful in handling many variables even in the presence of co-linearity and noise in the independent and dependent variables [36]. In this work, GA-PLS was utilised to choose the most informative molecular descriptors for the construction of the QSRR models. The chance of over-fitting was minimised by optimizing the number of latent variables (LVs). PLS models with a number of latent variables up to 5 were investigated and the optimum number of LVs in each model was selected by applying the first standard deviation rule [32, 37, 38].

2.3.5 Types of QSRR model

Two types of models can be generated in a QSRR study. The first is a global model, where a single model is built for the retention prediction of all compounds. The second is a local model, where a new model is derived for each new compound for which retention is to be predicted. The global modelling approach is popular in QSRR modelling because of its simplicity, but its major drawback is that the accuracy of prediction is generally low. On the other hand, local modelling where each compound has its own specific model, generally provides higher prediction accuracy. For both modelling approaches, a suitable training set is critical as this plays a major role in the prediction performance of the constructed QSRR model. Compounds in training sets can be selected either randomly or using targeted strategies to identify the best training set. The latter approach can be described as database "filtering" and this term is used throughout this thesis to describe the process of identifying the most appropriate compounds to be used in the training set. For example, filtering can be performed using the concept of structural similarity between the target analyte and the database compounds, based on the premise that a training set comprising structurally similar compounds would give a more accurate QSRR model than if the training set compounds had been selected randomly. Other parameters which can form the basis of filtering methods include analyte physico-chemical parameters (for example, log D which reflects the hydrophobicity), the nature of compounds (acids, bases, and neutrals), or the retention of compounds (retention time, retention factor). A more complex filtering approach using the second dominant interaction between compounds and stationary phase after hydrophobicity is another option. Regardless of which filtering method is applied, finding the final training set which provides the highest prediction accuracy is the ultimate goal.

2.3.6 Model validation

In QSRR modelling, a training set is used to build QSRR models using the most informative molecular descriptors, selected by the GA, and a test set is needed for validation [39-42]. Also, to evaluate the predictive ability of the constructed QSRR models, a separate external set is required [26, 43]. For this purpose, the measured chromatographic retention data of the test compounds were extracted and compared with their predicted retention data calculated from derived QSRR models. To generate test sets, a D-optimal algorithm was employed to split compounds in the dataset into a training set and a test set, respectively.

In this work, the coefficient of determination (R^2), the slope of the regression with no forced intercept, the mean absolute error (MAE) and the root-mean-square error of prediction (RMSEP) were utilised to evaluate model fitness and the predictive ability of the constructed QSRR models, with the requirement for the slope to be within the range of 0.85 to 1.15 [26,

34, 35]. The percentage root-mean-square error of prediction (RMSEP%) of retention time for the test set was measured to externally validate the accuracy of GA-PLS models generated from the training set.

MAE was defined as:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} \left| yi - \hat{y}i \right| \qquad 2.1$$

Where the yi and $\hat{y}i$ are, respectively, the experimental and predicted values of the response for the *i*-th compound in the dataset, and *n* is the number of compounds.

RMSEP was defined as:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}(yi(exp) - yi(pred))^2}{n}} \qquad 2.2$$

Where the *yi*(*exp*) and *yi*(*pred*) are, respectively, the experimental and predicted values of the response for the *i*-th compound in the dataset, and *n* is the number of compounds.

%RMSEP was defined as:

$$\%RMSEP = \sqrt{\frac{\sum_{i=1}^{n}\left(\frac{yi(exp) - yi(pred)}{yi(exp)}\right)^2}{n}} \times 100 \qquad 2.3$$

Where the *yi*(*exp*) and *yi*(*pred*) are the experimental and predicted retention times of the response for the *i*-th compound, and *n* is the number of compounds.

In Chapter 4, several filtering approaches were employed to generate training sets for the construction of QSRR models, and the predictive ability of the yielded models was evaluated by inspecting the Regression Error Characteristic (REC) curves obtained by plotting the prediction error range against the percentage of data points predicted within that range [44]. Furthermore, the overall performance of the above constructed models was further compared using the sum of ranking difference (SRD) approach where parameters for each model were compared to a series of reference values, and each model ranked according to how large was the difference between its parameters and the reference values [45, 46]. The rankings were also compared to a confidence interval generated by using randomly ranked numbers [45, 46]. More detail can be found in Chapter 4.

2.4 References

1. Wilson, N., M. Nelson, J. Dolan, L. Snyder, R. Wolcott, and P. Carr, *Column selectivity in reversed-phase liquid chromatography: I. A general quantitative relationship.* Journal of Chromatography A, 2002. **961**(2): p. 171-193.

2. LC Tan, PW Carr, and M. Abraham, *Study of retention in reversed-phase liquid chromatography using linear solvation energy relationships I. The stationary phase.* Journal of Chromatography A, 1996. **752**: p. 1-18.
3. *University of Minnesota - Boswell Research Group,* http://www.hplccolumns.org/database/index.php.
4. Hall, L.M., D.W. Hill, L.C. Menikarachchi, M.-H. Chen, L.H. Hall, and D.F. Grant, *Optimizing artificial neural network models for metabolomics and systems biology: an example using HPLC retention index data.* Bioanalysis, 2015. **7**(8): p. 939-955.
5. *MarvinSketch*. ChemAxon, (2016), chemaxon.com.
6. Halgren, T.A., *Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94.* Journal of Computational Chemistry, 1996. **17**(5-6): p. 490-519.
7. Halgren, T.A., *Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions.* Journal of Computational Chemistry, 1996. **17**(5-6): p. 520-552.
8. Halgren, T.A., *Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94.* Journal of Computational Chemistry, 1996. **17**(5-6): p. 553-586.
9. Halgren, T.A., *Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules.* Journal of Computational Chemistry, 1996. **17**(5-6): p. 616-641.
10. Halgren, T.A. and R.B. Nachbar, *Merck molecular force field. IV. Conformational energies and geometries for MMFF94.* Journal of Computational Chemistry, 1996. **17**(5-6): p. 587-615.
11. Vainio, M.J. and M.S. Johnson, *Generating conformer ensembles using a multiobjective genetic algorithm.* Journal of Chemical Information and Modeling, 2007. **47**(6): p. 2462-2474.
12. *MOPAC (2012)*. Stewart Computational Chemistry, Colorado Springs: CO, USA, OpenMOPAC.net.
13. Becke, A.D., *A new mixing of Hartree–Fock and local density-functional theories.* Journal of Chemical Physics, 1993. **98**(2): p. 1372-1377.
14. Hammer, B., L.B. Hansen, and J.K. Nørskov, *Improved adsorption energetics within density-functional theory using revised Perdew-Burke-Ernzerhof functionals.* Physical Review B, 1999. **59**(11): p. 7413.
15. Yang, W. and P.W. Ayers, *Density-functional theory*, in *Computational Medicinal Chemistry for Drug Discovery*. 2003, CRC Press. p. 103-132.
16. *in, Talete srl, Dragon 6.0 for Windows (Software For Molecular Descriptor Calculations);* http://www.talete.mi.it/ Talete, Milano, Italy.
17. *Matlab,in The Mathworks Inc., Natick, MA, USA, 2013.*
18. Katkova, E.V., I.V. Oferkin, and V.B. Sulimov, *Application of the PM7 quantum chemical semi-empirical method to the development of new urokinase inhibitors.* Vychisl. Metody Programm, 2014. **15**(2): p. 258-273.
19. Lee, C., W. Yang, and R.G. Parr, *Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density.* Physical review B, 1988. **37**(2): p. 785.
20. Stephens, P., F. Devlin, C. Chabalowski, and M.J. Frisch, *Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields.* The Journal of Physical Chemistry, 1994. **98**(45): p. 11623-11627.
21. Frisch, M.J., J.A. Pople, and J.S. Binkley, *Self-consistent molecular orbital methods 25. Supplementary functions for Gaussian basis sets.* Journal of Chemical Physics, 1984. **80**(7): p. 3265-3269.
22. Tomasi, J., B. Mennucci, and R. Cammi, *Quantum mechanical continuum solvation models.* Chemical Reviews, 2005. **105**(8): p. 2999-3094.

23. Tyteca, E., S.H. Park, R.A. Shellie, P.R. Haddad, and G. Desmet, *Computer-assisted multi-segment gradient optimization in ion chromatography.* Journal of Chromatography A, 2015. **1381**: p. 101-109.
24. Tyteca, E., M. Talebi, R. Amos, S.H. Park, M. Taraji, Y. Wen, R. Szucs, C.A. Pohl, J.W. Dolan, and P.R. Haddad, *Towards a chromatographic similarity index to establish localized quantitative structure-retention models for retention prediction: use of retention factor ratio.* Journal of Chromatography A, 2017. **1486**: p. 50-58.
25. Talebi, M., S.H. Park, M. Taraji, Y. Wen, R.I. Amos, P.R. Haddad, R. Shellie, R. Szucs, C. Pohl, and J.W. Dolan, *Retention time prediction based on molecular structure in pharmaceutical method development: A perspective.* LCGC North America, 2016. **34**(8): p. 550-558.
26. Wen, Y., M. Talebi, R.I. Amos, R. Szucs, J.W. Dolan, C.A. Pohl, and P.R. Haddad, *Retention prediction in reversed phase high performance liquid chromatography using quantitative structure-retention relationships applied to the Hydrophobic Subtraction Model.* Journal of Chromatography A, 2018. **1541**: p. 1-11.
27. Clementi, S., G. Cruciani, P. Fifi, D. Riganelli, R. Valigi, and G. Musumarra, *A new set of principal properties for heteroaromatics obtained by GRID.* Molecular Informatics, 1996. **15**(2): p. 108-120.
28. Cruciani, G., M. Pastor, and S. Clementi, *Handling information from 3D grid maps for QSAR studies*, in *Molecular modeling and prediction of bioactivity*. 2000, Springer. p. 73-81.
29. Cruciani, G., P. Crivori, P.-A. Carrupt, and B. Testa, *Molecular fields in quantitative structure–permeation relationships: the VolSurf approach.* Journal of Molecular Structure: THEOCHEM, 2000. **503**(1): p. 17-30.
30. Cruciani, G., M. Pastor, and W. Guba, *VolSurf: a new tool for the pharmacokinetic optimization of lead compounds.* European Journal of Pharmaceutical Sciences, 2000. **11**: p. S29-S39.
31. John, H., *Holland, Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. 1992, MIT Press, Cambridge, MA.
32. Leardi, R., *Application of genetic algorithm-PLS for feature selection in spectral data sets.* Journal of Chemometrics, 2000. **14**(5-6): p. 643-655.
33. Leardi, R. and A.L. Gonzalez, *Genetic algorithms applied to feature selection in PLS regression: how and when to use them.* Chemometrics and Intelligent Laboratory Systems , 1998. **41**(2): p. 195-207.
34. Taraji, M., P.R. Haddad, R.I. Amos, M. Talebi, R. Szucs, J.W. Dolan, and C.A. Pohl, *Prediction of retention in hydrophilic interaction liquid chromatography using solute molecular descriptors based on chemical structures.* Journal of Chromatography A, 2017. **1486**: p. 59-67.
35. Park, S.H., P.R. Haddad, M. Talebi, E. Tyteca, R.I. Amos, R. Szucs, J.W. Dolan, and C.A. Pohl, *Retention prediction of low molecular weight anions in ion chromatography based on quantitative structure-retention relationships applied to the linear solvent strength model.* Journal of Chromatography A, 2017. **1486**: p. 68-75.
36. Talebi, M., G. Schuster, R.A. Shellie, R. Szucs, and P.R. Haddad, *Performance comparison of partial least squares-related variable selection methods for quantitative structure retention relationships modelling of retention times in reversed-phase liquid chromatography.* Journal of Chromatography A, 2015. **1424**: p. 69-76.
37. Varmuza, K., P. Filzmoser, and M. Dehmer, *Multivariate linear QSPR/QSAR models: Rigorous evaluation of variable selection for PLS.* Computational and Structural Biotechnology Journal, 2013. **5**(6): p. e201302007.
38. Varmuza, K. and P. Filzmoser, *Introduction to multivariate statistical analysis in chemometrics*. 2016: CRC press.

39. Ghasemi, J. and S. Saaidpour, *QSRR prediction of the chromatographic retention behavior of painkiller drugs.* Journal of Chromatographic Science, 2009. **47**(2): p. 156-163.
40. Goryński, K., B. Bojko, A. Nowaczyk, A. Buciński, J. Pawliszyn, and R. Kaliszan, *Quantitative structure–retention relationships models for prediction of high performance liquid chromatography retention time of small molecules: endogenous metabolites and banned compounds.* Analytica Chimica Acta, 2013. **797**: p. 13-19.
41. Héberger, K., *Quantitative structure–(chromatographic) retention relationships.* Journal of Chromatography A, 2007. **1158**(1-2): p. 273-305.
42. Žuvela, P., J.J. Liu, K. Macur, and T. Baczek, *Molecular descriptor subset selection in theoretical peptide quantitative structure–retention relationship model development using nature-inspired optimization algorithms.* Analytical Chemistry, 2015. **87**(19): p. 9876-9883.
43. Taraji, M., P.R. Haddad, R.I. Amos, M. Talebi, R. Szucs, J.W. Dolan, and C.A. Pohl, *Use of dual-filtering to create training sets leading to improved accuracy in quantitative structure-retention relationships modelling for hydrophilic interaction liquid chromatographic systems.* Journal of Chromatography A, 2017. **1507**: p. 53-62.
44. *J. Bi, K.P. Bennett, Regression Error Characteristic Curves, Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003, pp. 43-50.*
45. Héberger, K., *Sum of ranking differences compares methods or models fairly.* Trends in Analytical Chemistry, 2010. **29**(1): p. 101-109.
46. Héberger, K. and K. Kollár-Hunek, *Sum of ranking differences for method discrimination and its validation: comparison of ranks with random numbers.* Journal of Chemometrics, 2011. **25**(4): p. 151-158.

3 Direct Prediction of Retention using Quantitative Structure-Retention Relationships in Reversed-Phase Liquid Chromatography

3.1 Introduction

Analytical method development (MD) is a key element of any pharmaceutical development program but it is often a time-consuming and labour-intensive process [1-4]. The workflow of systematic chromatographic method development contains two phases: *scoping* and *optimisation* [5-7]. As the primary phase of method development, *scoping* involves the selection of the preferred chromatographic technique, stationary phase, and broad composition of mobile phase. A subsequent phase, called *optimisation*, can then be performed to optimise and fine tune the selected chromatographic conditions by implementing experimental design approaches [5, 8]. An HPLC analysis method is developed to identify, quantify or purify compounds of interest, thus successful MD requires the experience and expertise of the chromatographer [9-11].

Computer-aided MD has been given intensive study as it can accelerate the MD process significantly if sufficiently accurate retention models exist [12, 13]. Nowadays, a wide variety of *in silico* tools have been used to speed up chromatographic MD based upon predicting retention from chemical structures [14, 15]. Commercial software such as Drylab (Molnár-Institute for Applied Chromatography, Berlin, Germany), ChromSword (ChromSword, Riga, Latvia) and ACD/ChromGenius (ACD/Labs, Toronto, Canada) have been utilised in chromatographic method development, optimisation, and validation [14, 15]. These software packages rely on experimental retention data and calculations for the chemical structures of compounds to predict retention under the same chromatographic conditions that have been used for compounds in the embedded database. In addition, when using these computer-aided MD software packages, the systematic experimentation that is undertaken also allows the implementation of Quality-by-Design (QbD) practices by providing tools to improve the robustness of the chosen chromatographic method and reduce labour and solvent consumption by limiting the required number of experiments [16, 17].

For the *scoping* phase of computer-aided MD a prediction error of up to 10% can be tolerated for the purpose of choosing some broad method parameters [4, 18, 19], including the most suitable chromatographic technique (such as reversed-phase [RPLC] [20, 21], hydrophilic interaction liquid chromatography [HILIC] [22], ion chromatography [IC] [8] or supercritical fluid chromatography [SFC] [23]), the stationary phase and mobile phase, *etc*. While for *optimisation*, prediction errors as low as one or two percent are needed to be able to find the optimal conditions for the separation of given compounds [24]. Quantitative Structure-Retention Relationship (QSRR) methodology has the potential to speed up the

scoping phase as exploratory experimentation can be replaced by retention prediction based solely on the chemical structures of molecules. Then, the best starting point for the *optimisation* phase of MD can be selected, after comparing retention prediction across a broad range of chromatographic techniques for a group of compounds [5, 25, 26]. The *optimisation* phase will always involve detailed experiments to measure retention accurately [27, 28].

QSRR provides a tool to generate more extensive information for retention phenomena including mechanism investigation, retention prediction and method development in chromatography [29, 30]. A QSRR model is usually created from a set of descriptors, either experimentally determined or theoretically computed from a symbolic representation of the molecules using commercial software tools [5, 30]. For some simple QSRR models, a limited number of pre-selected physico-chemical parameter descriptors are used. Examples can be found in ChromSword software where descriptors of the molecular volume and the energy of interaction with water are employed [11, 31]. Another software tool, ACD/ChromGenius uses more parameters such as log P, the log of the compound distribution coefficient (log D), polar surface area, molecular volume, molecular weight, molar refractivity and the number of hydrogen bond donor and acceptor sites on the molecule as descriptors to build QSRR models [5, 32].

As an alternative, a large pool of molecular descriptors can be generated using commercial software, such as Dragon [5, 8, 22, 33]. However, this should be followed by an appropriate variable selection strategy to extract the most relevant and informative descriptors for their subsequent use in QSRR modelling. A QSRR model with too few descriptors could be under-fitted and hence be insufficiently predictive, but a model with too many descriptors can increase the risk of over-fitting and introduce noise [1, 5]. Considering the large number of descriptors generated, implementing a suitable variable selection method, such as a genetic algorithm (GA), which is often combined with multiple linear regression (MLR) or partial least squares regression (PLS), becomes necessary to exclude noise from the model and reduce the risk of over-fitting and chance correlation [12, 15]. PLS regression is particularly useful in the presence of co-linear, redundant and noisy variables, and in handling databases with a high number of variables compared to the number of sample compounds [5, 34]. It has been shown that the performance of PLS modelling can be improved significantly by applying a suitable feature selection method [34, 35]. As reported by P Žuvela *et al.* [36], a combination of GA and PLS was best for selecting the most important and relevant descriptors compared to other optimisation algorithms in terms of computational cost, accuracy and robustness of the constructed QSRR models.

QSRR models either can be built using the whole dataset (all compounds except the target compound), or a group of compounds from that dataset, as the training set [18]. The use of the whole dataset as training set is popular in QSRR modelling, but the drawback of this approach is that, most of the time, the accuracy of prediction is unsatisfactory [18, 37]. The use of a specific training set of compounds can improve the accuracy of prediction results as the concept of similarity is often used to form training sets. It has been shown that smaller, more similar training sets to the target compound lead to greater prediction accuracy [18, 38, 39]. Similar compounds to the target can be selected using some criterion, such as the similarity of chemical structures between molecules, the proximity of physico-chemical properties, or the retention parameters of the compounds of interest.

In the present study, the ratio of retention factor was used as a chromatographic similarity filter to yield training sets for the construction of QSRR models. Furthermore, the Tanimoto approach, which can be seen the gold standard in computing fingerprint-based similarity, was also investigated and compared as a filter for building training sets for QSRR modelling. In addition, in order to find a chromatographic similarity index which is comparable with the *k*-ratio filter, the representative molecular descriptors log D and log P were also explored as filters for the training sets. Finally, the effectiveness of a dual filter that uses Tanimoto or log D as the primary, and the *k*-ratio as the secondary filter was also evaluated.

3.2 Materials and methods

3.2.1 Database

Three retention datasets used in the present study have been described previously in section 2.1, Chapter 2, including the names of compounds (Tables 2.1, 2.3 and 2.4 in Chapter 2), the retention information (Tables 2.5, 2.6 and 2.7 in Chapter 2), the characteristics of the columns (section 2.1 and Table 2.1 in Chapter 2), and the chromatographic conditions used. Retention predictions were performed for compounds in Dataset 1 on ten columns (column number 1 to 10), compounds in Dataset 2 on five columns (column number 11 to 15), and compounds in Dataset 3 on a Zorbax Eclipse Plus C_{18} column (column number 16).

3.2.2 Calculation of molecular descriptors

In this study, Dragon descriptors were calculated and employed. The Dragon 6.0 software [40] is able to calculate in excess of 4000 molecular descriptors, consisting of constitutional, topological, geometrical, electrostatic, physical, and quantum chemical descriptors [41]. The calculations of molecular descriptors were performed as detailed in section 2.3.2, Chapter 2. The resulting descriptor sets were used to build predictive models for the experimental

chromatographic retention data. Finally, 1448 descriptors were calculated and exported for each compound in the three datasets.

3.2.3 Similarity ranking

In the present study, several filters for similarity ranking were applied to the datasets to generate suitable training sets. Each compound in each dataset was subsequently utilised as a 'target compound'. Its retention time was then predicted using models made up from a subset of the other compounds in the dataset, by treating the other compounds with various filters. Filters included Tanimoto (based on the similarity of chemical structure), physico-chemical properties such as Log D and Log P, the chromatographic similarity reflected by the ratio of retention factor (*k*-ratio), and a dual filter, using Tanimoto or log D as the primary, and *k*-ratio as the secondary filter.

Tanimoto similarity: in the first modelling approach, filtering was performed based on the Tanimoto Similarity (TS) index. One compound (the target compound) was left out of the dataset and the rest sorted based on their pairwise TS index in relation to the target. Then the top ten compounds were used as a training set to derive a separate QSRR model for retention factors. If ten compounds could not be found in the training set, the target compound was not modelled. The derived models were then used to predict the retention factor for the target and this process was repeated for each of the compounds in the dataset. The TS-values were calculated using JChem for Excel (ChemAxon, Budapest, Hungary).

Physico-chemical parameter similarity (represented by log D and log P): some representative descriptors such as log D and log P can be used as filters to derive training sets. In the log D approach, filtering was performed based on the log D values. One compound (the target compound) was left out of the dataset with the rest being sorted based on the difference of log D in relation to the target. The largest difference allowed for filtering was 0.2. Similarly, for the log P approach, compounds in the training set were selected based on the ratio of log P (with the ratio always > 1) to the target compound and the largest ratio allowed was 1.2. Finally, the selected compounds were used as a training set to derive a QSRR model for the retention factor of the target compound. The minimum number of compounds in the training set was five, if five compounds below the respective cut-off similarities could not be found, the target compound was not modelled. The log D and log P values at the pH of the RPLC mobile phase (pH = 2.8) were calculated using InstantJChem (ChemAxon).

Chromatographic similarity (represented by *k*-ratio index): it is well known that the design of a practically useful similarity index should in fact correspond to the chromatographic similarity between compounds in order to establish accurate predictive QSRR models [5].

Therefore, the ratio of retention factors was also considered as a filter in the present study although it cannot be applied in practice. For *k*-ratio filtering, the compounds in the database were ranked according to the ratio of the compound's retention factor *k* with the *k*-value of the reference compound (with *k*-ratio always > 1). The training set was then built using a certain *k*-ratio threshold (*k*-ratio < 1.5 in this study) to construct predictive QSRR models. The minimum number of compounds in the training set was five.

Dual filter: the *k*-ratio approach cannot be applied in practice because the retention of the target compound is unknown. The proposed *k*-ratio filter is therefore useful only as a benchmark. However, the retention of the compound could be used as a secondary filter after the initial application of a Tanimoto or log D (or log P) filter. Therefore, a secondary *k*-ratio filter was applied to datasets that had been determined using Tanimoto, log D, or log P as the primary filter. The rationale is to first select a training set based on a primary filter (such as Tanimoto or log D) and to then scrutinise the retention times of the training set compounds and to remove any compounds which have very diverse retention times.

3.2.4 QSRR modelling

In the present study, the QSRR models were obtained *via* a PLS regression in combination with a GA as the variable selection method [5, 35]. The parameters of the GA were detailed in section 2.3.3, Chapter 2. The similarity ranking, descriptor selection, and QSRR modelling were performed in an automated fashion using Matlab software. To enable this, the original GA-PLS Matlab routines from Leardi were modified [34].

3.2.5 Statistics

The coefficient of determination (R^2), the slope of the regression with no forced intercept and the mean absolute prediction errors (MAE) were used to evaluate model fitness (as detailed in section 2.3.6, Chapter 2) with the requirement for the slope to be within the range of 0.85 to 1.15 [42]. The correlation coefficient R^2 between the predicted and the experimental retention factors was calculated by constructing the corresponding scatter plot and performing a linear regression in Excel.

3.3 Results and discussion

3.3.1 Role of chromatographic similarity (k-ratio filter)

A fundamental premise of this study is that basing QSRR models on those compounds in the retention database that are chromatographically similar to the target compound will yield more accurate predictions than using the entire database to derive the models [5, 18, 35]. To illustrate this point, for each dataset, QSRR modelling was performed based only on those compounds in the dataset where a training set could be compiled that showed a retention factor

k-ratio < 1.5 to the target compound. Retention factor prediction was therefore performed for 90 compounds in Dataset 1 on ten columns (number 1 to number 10), 87 compounds in Dataset 2 on five columns (column number 11 to number 15), and 112 compounds in Dataset 3 on one column (column number 16).

Figure 3.1 shows the correlation between predicted and observed retention factors for the 90 compounds in Dataset 1 on ten columns. The QSRR modelling shows very high correlations between predicted and observed retention factors in all cases, with slopes ranging from 1.054 (Figure 3.1j) to 1.1019 (Figure 3.1i), and a correlation coefficient (R^2) range from 0.970 (Figure 3.1g) to 0.992 (Figure 3.1j), in addition, very small prediction errors were obtained for compounds in Dataset 1 on all ten columns, with MAE values from 0.099 (Figure 3.1e) to 0.565 (Figure 3.1a). The same trend was observed for the 87 compounds in Dataset 2 on all five columns as can be seen in Figure 3.2, with MAEs between 0.248 (Figure 3.2d) and 0.519 (Figure 3.2a). Similarly, data points are very close to the trend line with a range of correlation coefficients (R^2) from 0.961 (Figure 3.2b) to 0.974 (Figure 3.2d). The same approach resulted in a higher prediction error (MAE = 1.805, Figure 3.3) for the 112 compounds in Dataset 3 compared to previous datasets. The reason could be the much larger retention window (maximum retention factor > 95) for compounds on column number 16 compared to the retention windows for compounds in other two datasets. Also, as can be seen from Figure 3.3, especially for early eluted compounds, a very good correlation was observed (R^2 = 0.992), while the data points are more scattered for the more retained compounds (compounds with retention factor greater than 60).

It is worth noting that the retention factor filter cannot be applied in practice because the retention times of the target compounds are unknown. Therefore, an alternative filtering tool is required which enables the formation of a training set that contains the same compounds that were clustered into a training set by the retention factor filter. In this study, the use of a structural similarity index (Tanimoto Similarity Index) has been investigated in order to evaluate whether it can be used to attain similar prediction accuracy to that achieved by *k*-ratio filtering. Also, physico-chemical parameters (log D and log P) and a dual filter approach were also investigated.

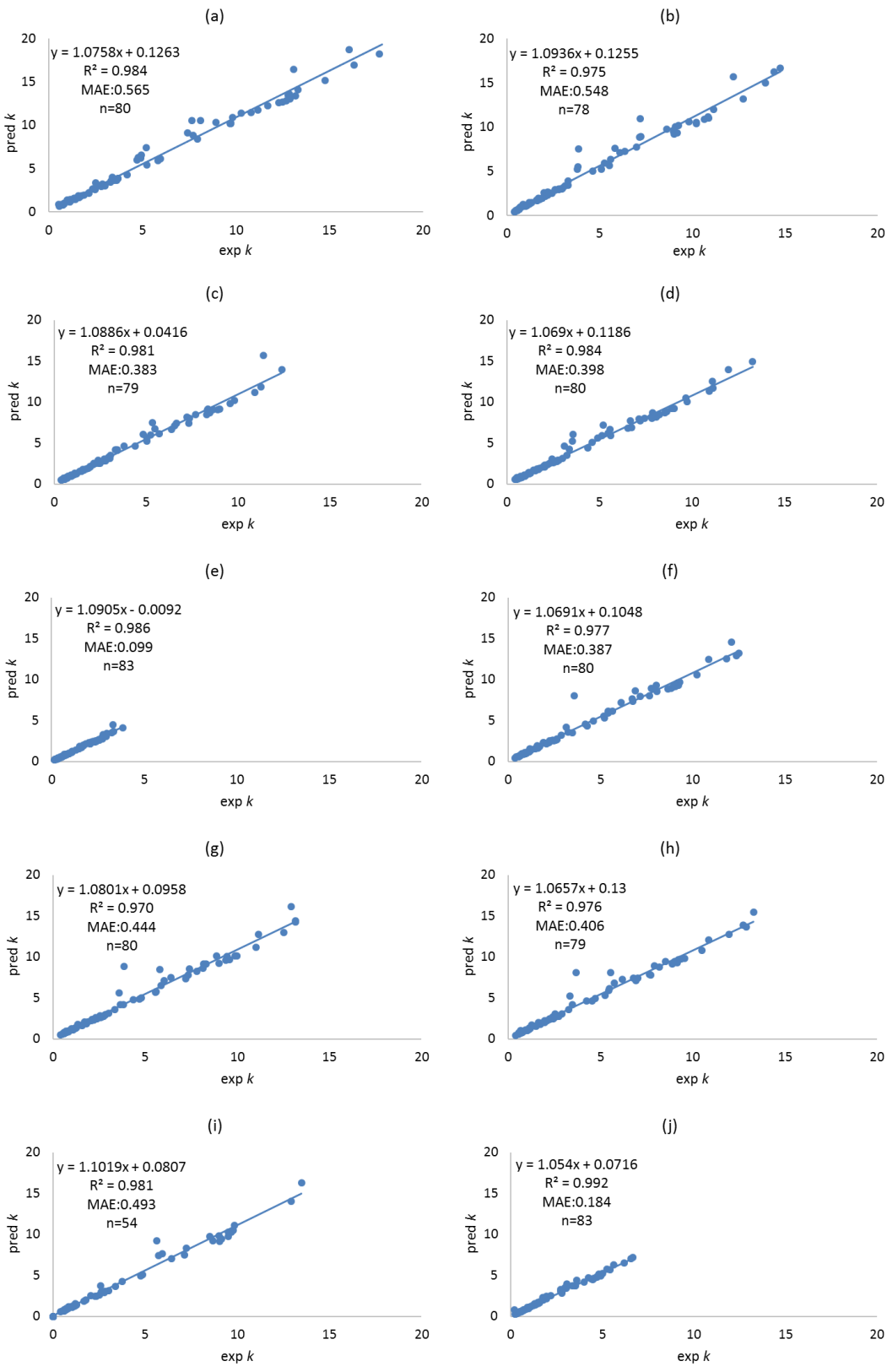


Figure 3.1. Predicted retention factors (*k*) of 90 compounds (Dataset 1) on ten columns using the *k*-ratio 1.5 filter. Numbering of the graphs from (a) to (j) represents the ten columns (number 1 to number 10).

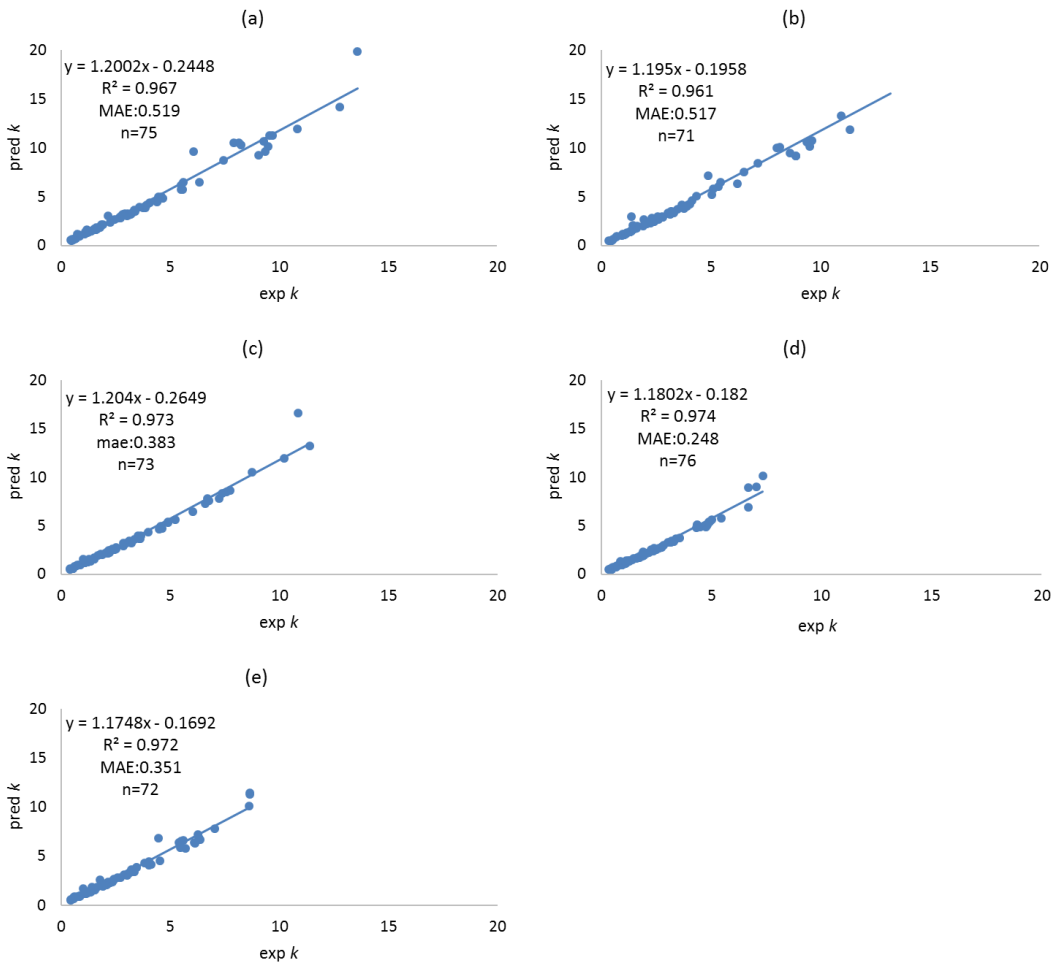


Figure 3.2. Predicted retention factors (*k*) of 87 compounds (Dataset 2) on five compounds using the *k*-ratio 1.5 filter. Numbering of the graphs from (a) to (e) represents the five columns (number 11 to number 15) utilised for Dataset 2.

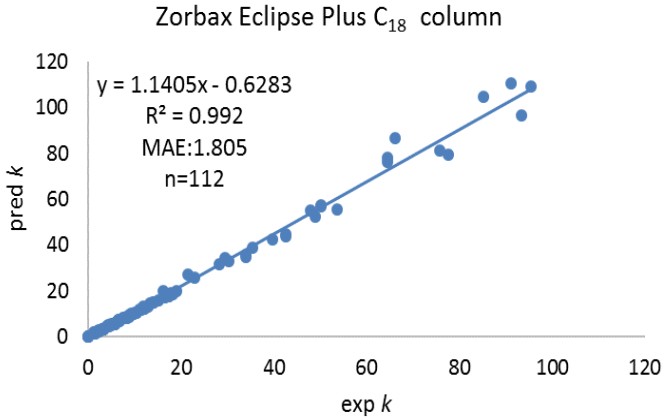


Figure 3.3. Predicted retention factors (*k*) of 112 compounds (Dataset 3) on Zorbax Eclipse Plus C_{18} (column number 16) using the *k*-ratio 1.5 filter.

3.3.2 QSRR modelling using Tanimoto similarity

The Tanimoto coefficient as a measure of molecular similarity was used to carry out TS searching in this study [32, 43]. When TS searching was employed, using the top 10 ranked similar compounds, the resulting models gave unacceptable prediction errors. Specifically, the Tanimoto approach resulted in a range of MAE from 0.687 (Figure 3.4e) to 3.732 (Figure 3.4a) for Dataset 1, multiplying the *k*-ratio errors ten-fold (Figure 3.1). The prediction error of 0.687 (Figure 3.4e) seems acceptable but it is worth pointing out that the retention window is quite narrow with values all less than 5. Meanwhile, the maximum value of the correlation coefficient was only 0.876 (Figure 3.4i), which is far less than expected. Almost 5-fold higher prediction errors were observed for the 87 compounds in Dataset 2 (column number 11 to 15) using the Tanimoto approach compared to the *k*-ratio approach (MAE values can be seen in Figure 3.5). As expected, a larger error of retention prediction (MAE = 15.859) resulted for Dataset 3 which contains many highly retained compounds on column number 16 (Figure 3.6). Large errors of prediction were obtained even for early eluted compounds using Tanimoto searching, indicating that the selected top ten compounds in the training set were not similar enough to the target compound. The high prediction errors obtained using the Tanimoto approach suggested the need to screen compounds into training sets where more similar compounds are kept and dissimilar compounds excluded. In other studies, it has been found that Tanimoto similarities of 0.7 and higher are necessary for accurate retention prediction [8, 44]. The training sets for this study had some compounds of TS 0.5 and lower. It is expected that an adequate threshold of similarity offering training sets with high enough average TS-values could significantly improve the accuracy of prediction in QSRR modelling.

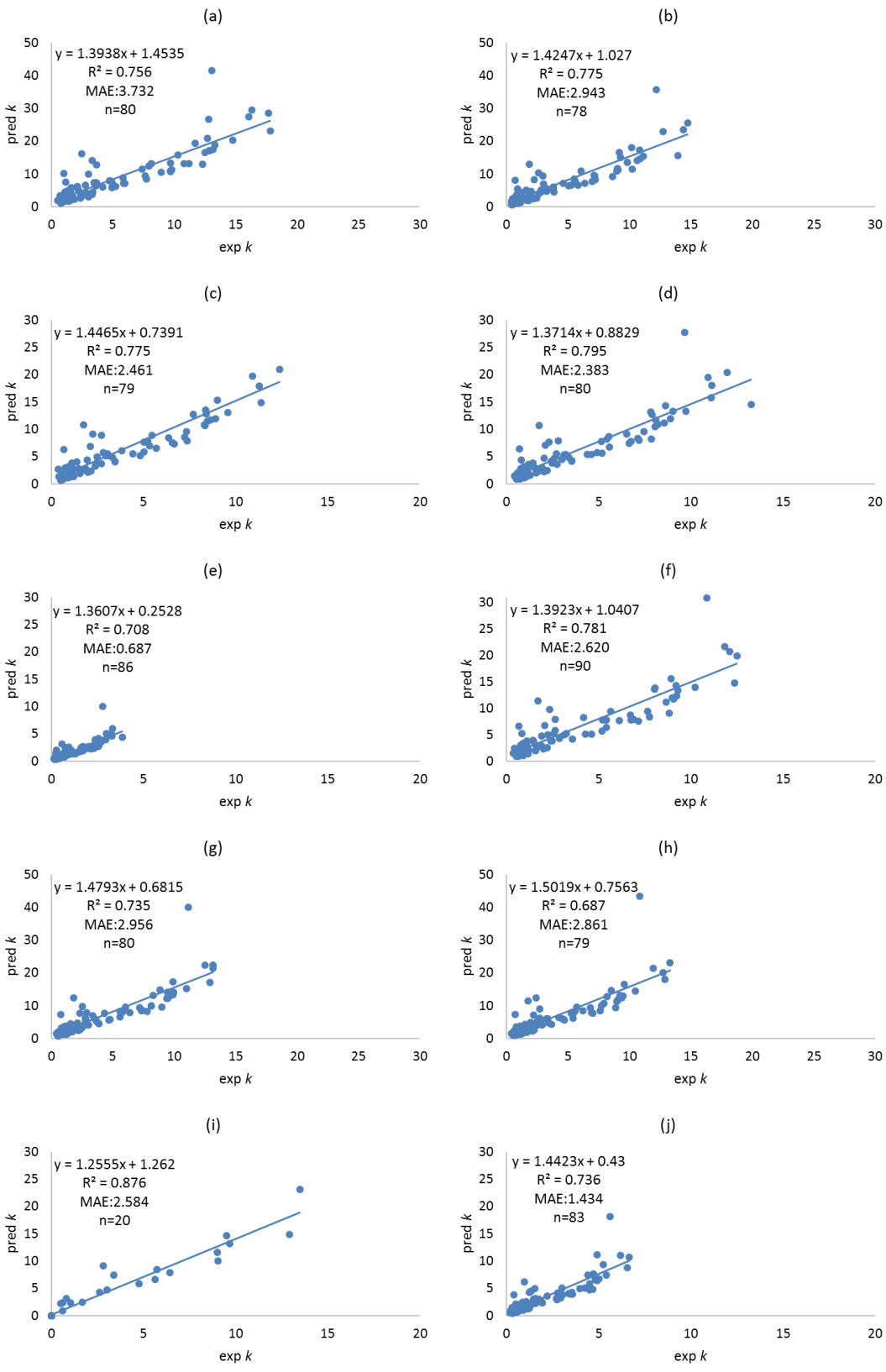


Figure 3.4. Predicted retention factors (*k*) of 90 compounds (Dataset 1) on ten columns using the Tanimoto filter. Numbering of the graphs from (a) to (j) represents the ten columns (number 1 to number 10).

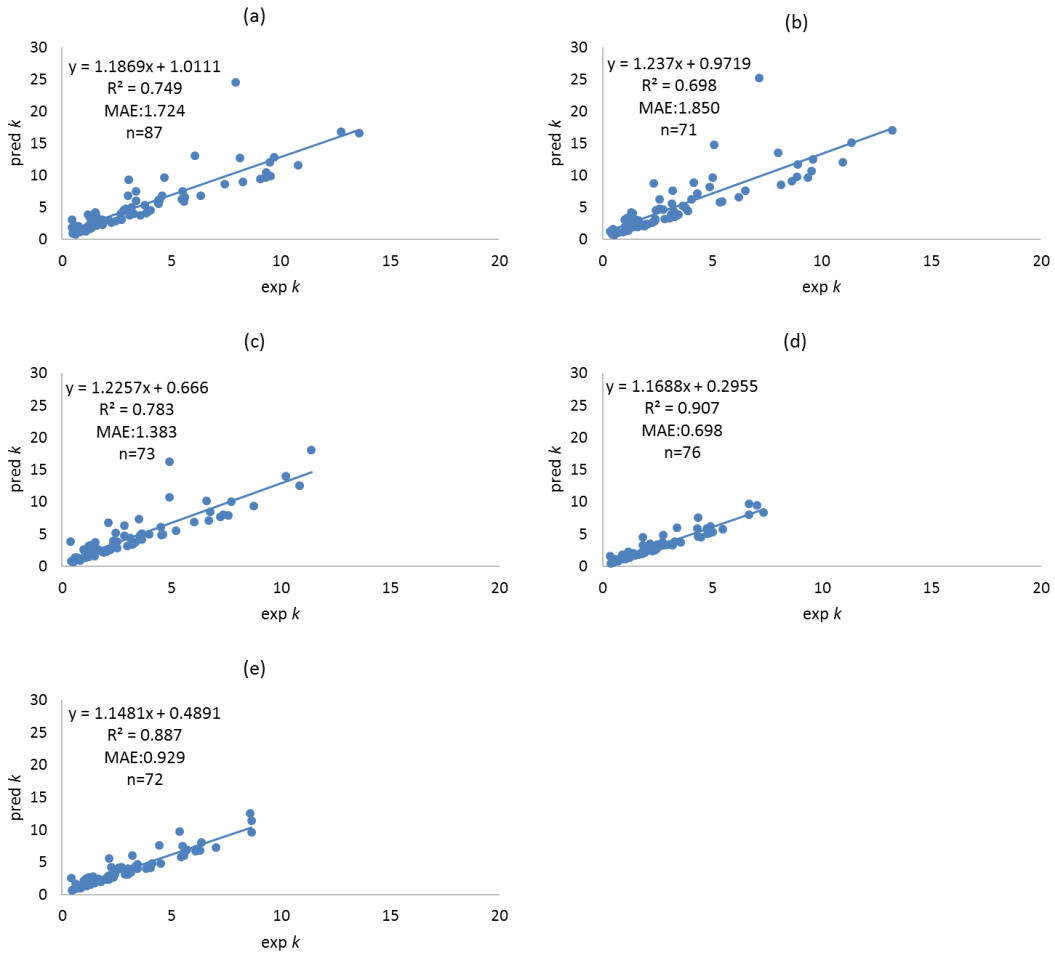


Figure 3.5. Predicted retention factors (*k*) of 87 compounds (Dataset 2) on five compounds using the Tanimoto filter. Numbering of the graphs from (a) to (e) represents the five columns (number 11 to number 15) utilised for Dataset 2.

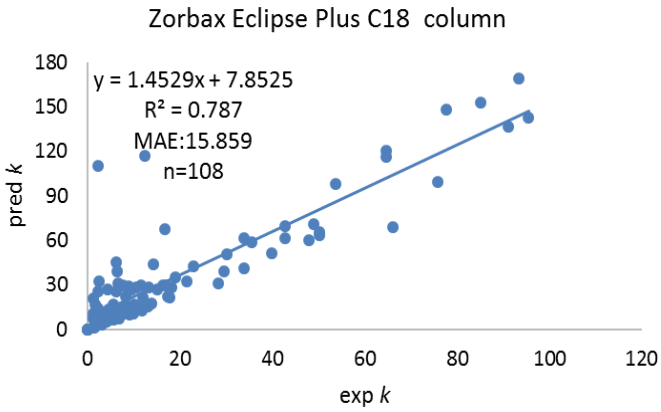


Figure 3.6. Predicted retention factors (*k*) of 112 compounds (Dataset 3) on Zorbax Eclipse Plus C_{18} (column number 16) using the Tanimoto filter.

3.3.3 QSRR modelling using Physico-chemical parameter similarity (represented by log D and log P)

Some important physico-chemical parameters – log D and log P were also investigated in this study as they represent the significant contribution of the hydrophobic interaction to retention in RPLC. As a further study, the absolute difference in log D between each compound in the dataset and the target compound was used to rank the compounds with a cut-off 0.2. For the log P filter, the only difference was in the selection of the threshold, instead of using the absolute difference of 0.2, a ratio of 1.2 was employed while other conditions were kept as same as for the log D approach. The log D approach was applied to compounds in all three datasets and columns, but it was not performed on the number 9 column in Dataset 1 as the retention data for some compounds on that column were not available (Table 2.5 in Chapter 2). The log P filter was only performed for compounds in Dataset 2 on columns 11 and 15. Figures 3.7 – 3. 10 show the correlations of measured and predicted retention factors using the log D and log P approaches, respectively.

As can be seen from Figures 3.7 – 3.9, much smaller retention prediction errors were obtained for compounds in each dataset compared to the MAE generated using the Tanimoto approach, but prediction errors were still larger than the *k*-ratio approach, as expected. The maximum and minimum MAE values of 2.538 (Figure 3.7a) and 0.502 (Figure 3.7e) were observed for 90 compounds in Dataset 1, both smaller than for the Tanimoto approach models generated using the same compounds and columns (MAE of 3.732 and 0.687, respectively). Similarly, the log D filter also resulted in much smaller MAEs for compounds in Dataset 2 on five columns compared to the models from the Tanimoto filter employing the same compounds and columns (Figure 3.8 and Figure 3.5). Again, although scattered data points for 112 compounds in Dataset 3 were still observed using the log D approach on column number 16, the correlation coefficient (R^2 = 0.820) and MAE (11.349) were both much improved compared to the Tanimoto approach (R^2 = 0.787 and MAE = 15.859, respectively).

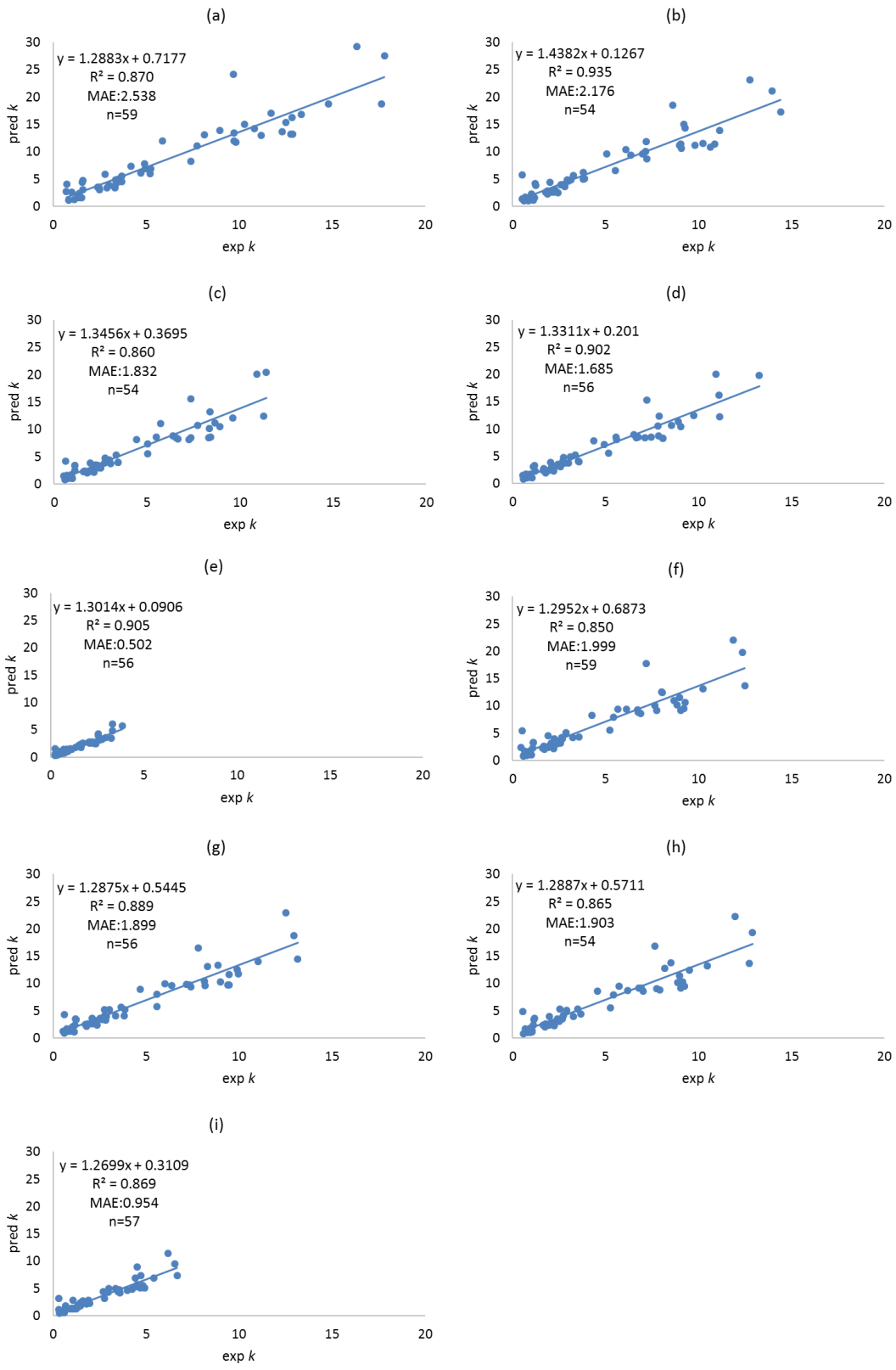


Figure 3.7. Predicted retention factors (*k*) of 90 compounds (Dataset 1) on nine columns using the log D filter. Numbering of the graphs from (a) to (i) represents the nine columns (number 1 to number 8 and column number 10).

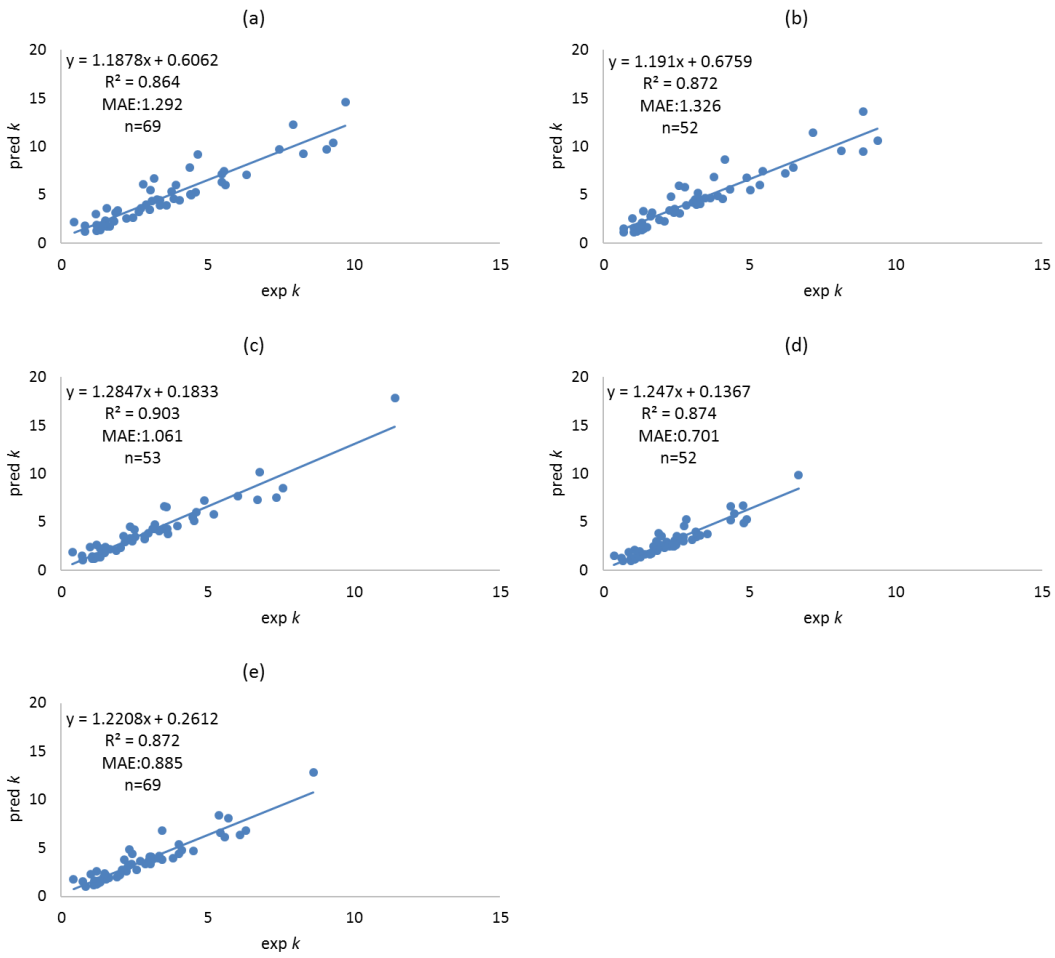


Figure 3.8. Predicted retention factors (*k*) of 87 compounds (Dataset 2) on five compounds using the log D filter. Numbering of the graphs from (a) to (e) represents the five columns (number 11 to number 15) utilised for Dataset 2.

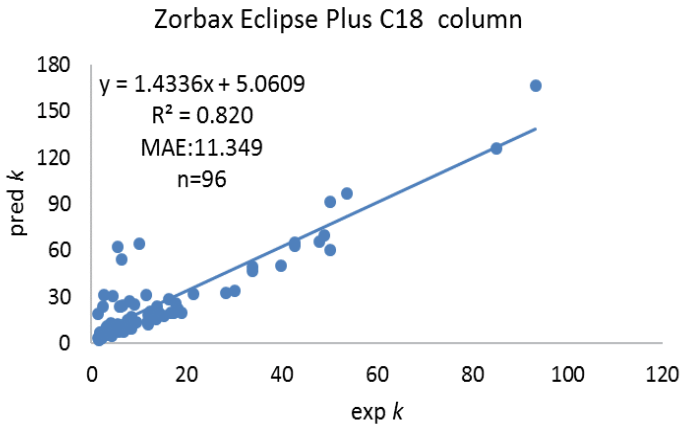


Figure 3.9. Predicted retention factors (*k*) of 112 compounds (Dataset 3) on Zorbax Eclipse Plus C_{18} (column number 16) using the log D filter.

In terms of the log P approach, it was evaluated only for the 87 compounds in Dataset 2 on columns 11 and 15. The prediction errors were larger (MAE of 2.549 for column 11 and 1.556 for column 15, Figure 3.10) than MAE values for the log D approach (MAE of 1.292

and 0.885, respectively), and even larger than for the Tanimoto approach (MAE of 1.724 and 0.929, respectively). Therefore, it can be assumed that log P does not explain the physico-chemical properties of the compounds that are most involved in chromatographic retention.

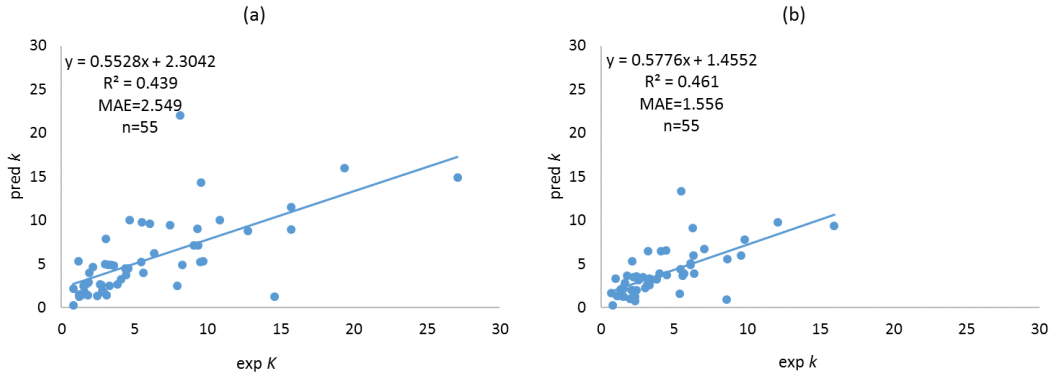


Figure 3.10. Predicted retention factors (*k*) of 87 compounds (Dataset 2) on two columns using the log P filter. Numbering of representative columns in graph: (a), Zorbax SB-C_{18} (column number 11); (b), Zorbax C_8 (column number 15).

It is to be noted that very large prediction errors were observed for the 112 compounds in Dataset 3 no matter which approach was applied except when the *k*-ratio was used as a filter. Dataset 3 consisted of 112 compounds with a large range in *k*-values (varying between 0.5 and 95.6). Also, Dataset 3 was sparsely populated for *k*-values higher than 20. However, those compounds are unlikely to be used in practical chromatographic method development. For Dataset 3, again only the filter using the *k*-ratio resulted in acceptable retention prediction.

The log D filter resulted in much better results and there is a clear improvement compared to the Tanimoto QSRR modelling. But since the errors are larger than 10%, which was the cut-off for scoping purposes, the log D filter still cannot be used in practice. In this study log D and log P descriptors were selected as the filters because they can reflect the main hydrophobicity interaction which has been recognised as the primary contributor to retention in RPLC. Using only log D or log P as the single filter did not yield acceptable retention prediction, and this is not surprising as in fact there is not a strong linear correlation between log D (or log P) and their retention factors. As can be seen from Figure 3.11 and Figure 3.12 as the log D (or log P) values increase, the corresponding retention factors also increase but the data points are very scattered from the trendline, and correlation coefficients are very small in all cases.

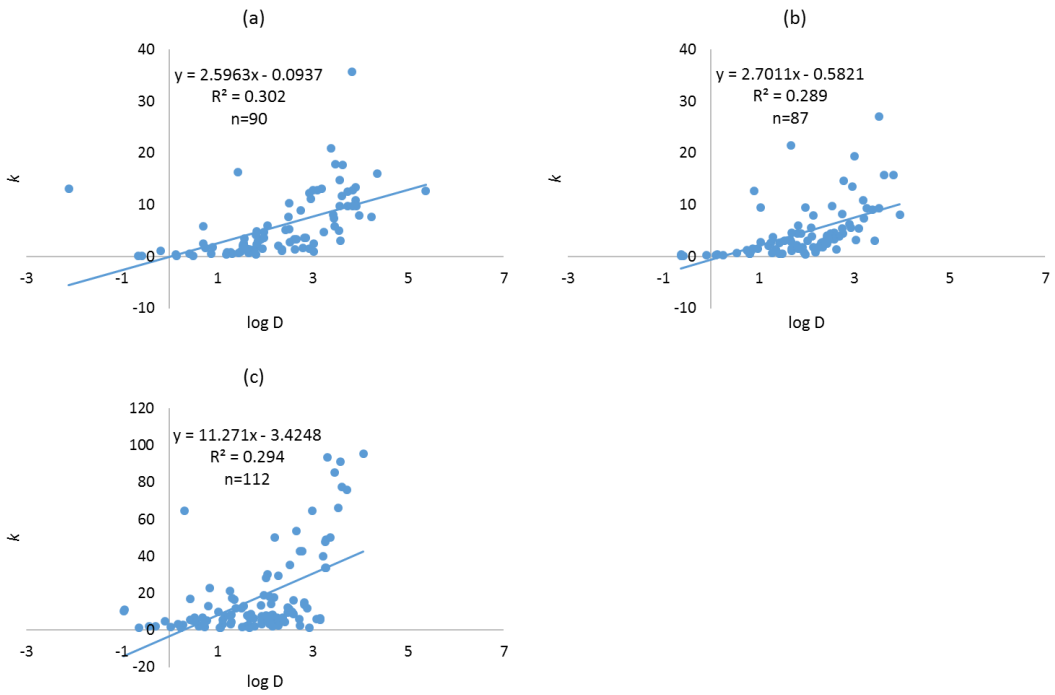


Figure 3.11. Correlation between values of the log D descriptor and retention factor (*k*) for: (a) 90 compounds in Dataset 1 and column number 1; (b) 87 compounds in Dataset 2 and column number 11, and (c) 112 compounds in Dataset 3 on column number 16.

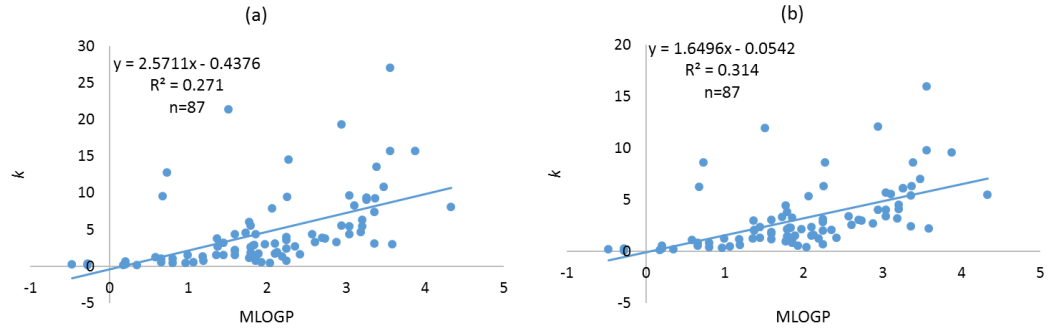


Figure 3.12. Correlation between values of the log P descriptor and retention factor (*k*) for 87 compounds in Dataset 2 on (a) Zorbax SB-C_{18} column (column number 11) and (b) Zorbax C_8 column (column number 15).

3.3.4 QSRR modelling using dual filter

The previous section has shown the superior predictive ability of the QSRR models when a *k*-ratio filter is used to build the training set. As stated previously, this approach cannot be applied in practice but might be used as a secondary filter after the initial application of the primary filter. The workflow is to select a training set based on a primary filter first and then scrutinise the retention factors of the training set compounds and remove any compounds which have very diverse retention factors. To investigate this approach, a secondary *k*-ratio filter was applied to datasets that had been determined using Tanimoto and log D as the primary filter. This dual filter approach was examined on the compounds in three datasets and the corresponding columns.

A *k*-ratio filter with a cut-off of 1.5 was applied to the compounds in the dataset after using Tanimoto or log D as the primary filter. Instead of applying the *k*-ratio filter to the target compound, the most similar compound in the training set identified from the Tanimoto or log D filter was used as a surrogate for the target compound, then the secondary filter was used to build the final training set. In the present study, the *k*-ratio secondary filter was tested on Datasets 1, 2, and 3, where Tanimoto and log D was used as the primary filter. As can be seen from Figure 3.13 and Figure 3.14, none of these dual filters resulted in improved predictions compared to the use of a single, primary filter (Tanimoto or log D). For Tanimoto – *k*-ratio dual filter, prediction errors (MAE) of 3.863 and 2.754 were observed for the 90 compounds in Dataset 1 on the columns 1 and 6, quite similar to the prediction errors yielded using Tanimoto as the single filter (MAE of 3.732 and 2.620, Figure 3.4a and 3.4f). Unfortunately, the dual filter resulted in MAE values of 3.135 and 1.944 for the 87 compounds in Dataset 2, which were much higher than the MAEs of 1.724 and 0.929 (Figure 3.5a and 3.5e) that resulted when only Tanimoto filtering was used for the same columns (column numbers 11 and 15). Similarly, for Dataset 3, a MAE of 15.211 was obtained after using the dual filter which is very close to the prediction error of MAE = 15.859 found using the Tanimoto filter alone (Figure 3.6).

For the second dual filter, log D combined with the *k*-ratio filter, much higher prediction errors and poorer correlations were observed for all three datasets. Specifically, the log D – *k*-ratio dual filter resulted in much higher prediction errors (MAE = 3.221 and 2.405, respectively) than MAE of 2.538 and 1.999 (Figure 3.7a and 3.7f) obtained using the log D single filter for Dataset 1 on columns 1 and 6. For the compounds in Dataset 2, larger predictions (MAE of 1.644 and 0.921) were yielded again compared to the log D approach (MAE of 1.292 and 0.885, Figure 3.8a and 3.8e). Similarly, a MAE of 9.474 was obtained after using *k*-ratio as the secondary filter for 112 compounds in Dataset 3, comparable to the errors using the log D filter only (MAE of 11.349, Figure 3.9).

These results are perhaps not surprising because the primary filter used here, Tanimoto for example, still stands on the similarity of chemical structure, which has proved to be inadequate in reflecting the chromatographic similarity for the more complex separation modes of RPLC. But, this situation is expected to improve when the dataset contains a large number of similar compounds [8, 44].

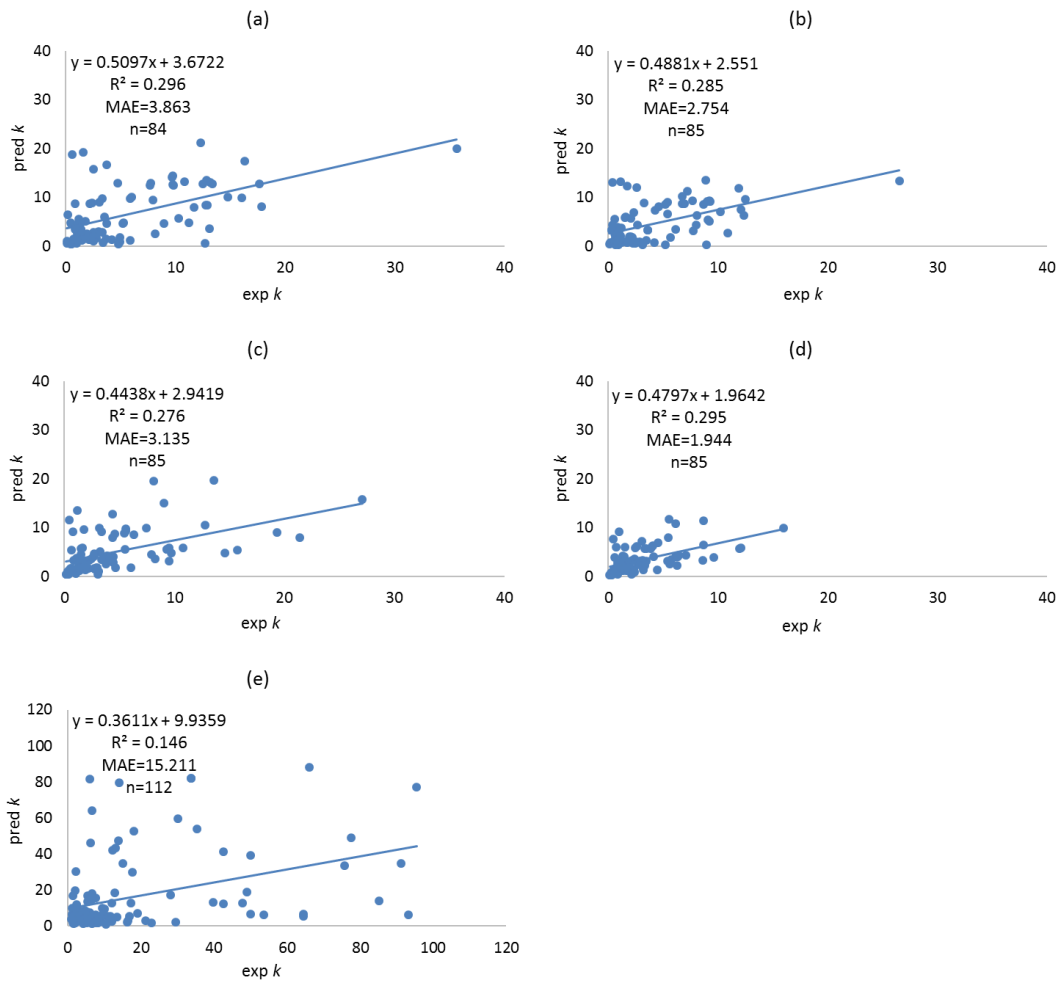


Figure 3.13. Predicted retention factors (*k*) of compounds using the dual filter: Tanimoto combined with *k*-ratio for (a), 90 compounds 1 in dataset and GL Inertsil ODS-3 (column number 1); (b), 90 compounds in Dataset 1 and HP Eclipse XDB-C_{18} (column number 6); (c) 87 compounds in Dataset 2 and Zorbax SB-C_{18} (column number 11); (d) Zorbax C_8 (column number 15) and (e), 112 compounds in Dataset 3 and Zorbax Eclipse Plus C_{18} (column number 16).

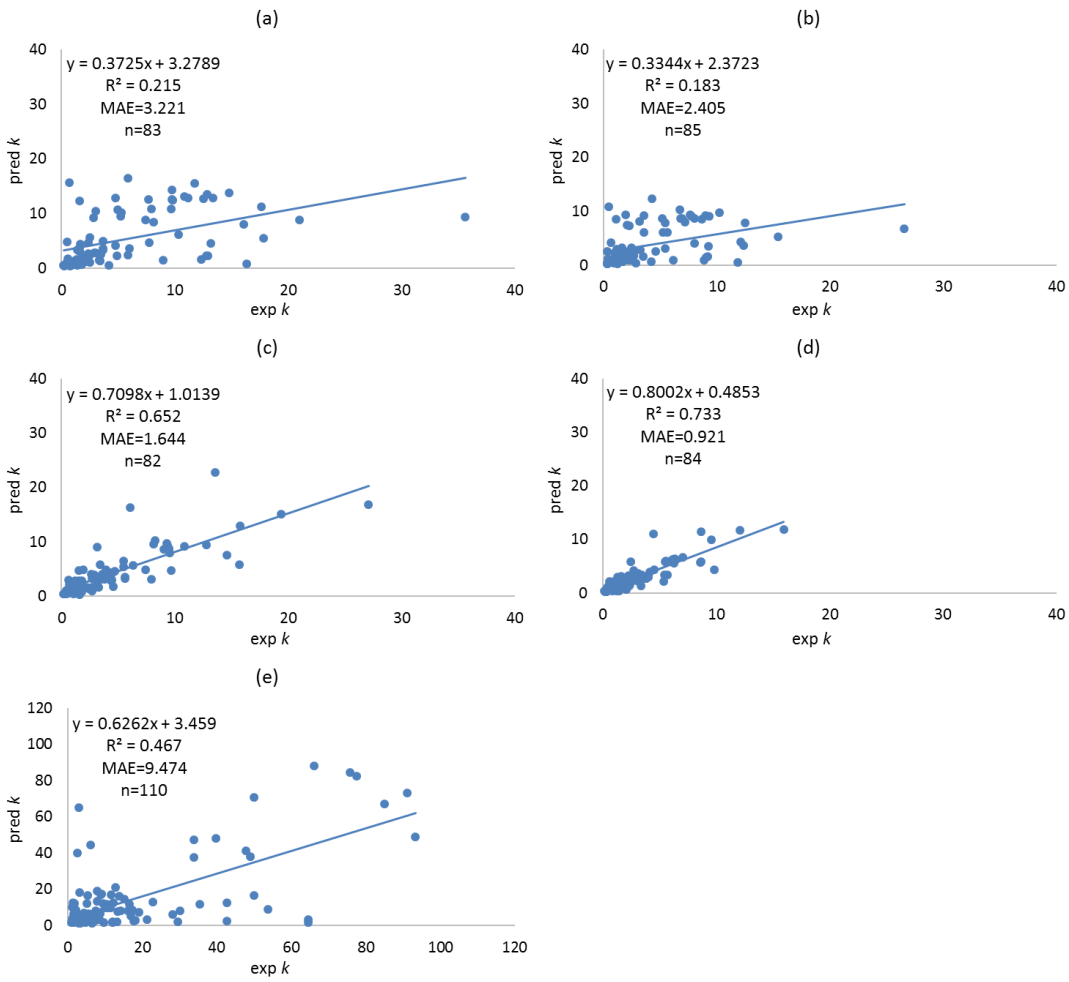


Figure 3.14. Predicted retention factors (*k*) of compounds using the dual filter: log D combined with *k*-ratio for (a), 90 compounds 1 in dataset and GL Inertsil ODS-3 (column number 1); (b), 90 compounds in Dataset 1 and HP Eclipse XDB-C_{18} (column number 6); (c) 87 compounds in Dataset 2 and Zorbax SB-C_{18} (column number 11); (d) Zorbax C_8 (column number 15) and (e), 112 compounds in Dataset 3 and Zorbax Eclipse Plus C_{18} (column number 16).

3.3.5 The importance of high Tanimoto score for QSRR modelling

It is expected that much better retention prediction will be obtained using structurally similar compounds to the target when building the training set. However, in this study when the Tanimoto top ten approach was applied, poor correlations between experimental and predicted retention were observed for all three datasets. Here, the training set was built using the top ten compounds in the dataset based on their Tanimoto similarity score to the target compound, which means there is a great possibility that some dissimilar compounds were also included. This speculation was confirmed by analysing the Tanimoto similarity scores of each training set (top ten similar compounds to the target). For the 90 compounds in Dataset 1, the average TS-values for the training sets identified for each target compound varied between 0.089 and 0.648 (an overall average of 0.432). For the 87 compounds in Dataset 2, slightly higher average TS-values were observed with a range of 0.108 to 0.719 (average of 0.494).

Very similar average TS-values for 112 compounds in Dataset 3 were obtained with a range between 0.186 and 0.715 (average of 0.434). The lower TS-values for some training sets indicated that many dissimilar compounds to the target had been selected to construct the training sets and QSRR models. Thus, the retention of the target compound was often modelled and predicted using a number of dissimilar compounds. One solution that can be considered is to use a threshold of Tanimoto similarity score to screen compounds, which would allow training sets for each target to be built based only on similar compounds. Research on the retention prediction in hydrophilic interaction liquid chromatography [HILIC] has shown a general trend towards smaller prediction errors at high TS-values (TS > 0.8), suggesting that a training set of sufficient similarity to the target compound should produce low prediction errors [8, 18, 22]. In present study, the maximum score of the average TS-values is only 0.719, showing that these datasets do not contain sufficient numbers of high-similarity compounds to provide training sets in which all compounds have TS > 0.8. This limitation of the database can be solved by including more diverse compounds so that the probability of finding a sufficient number of highly similar compounds is much higher.

3.4 Conclusions

In the present study, modelling strategies for retention prediction were investigated and evaluated in QSRR modelling. Strategies for compound filtering were employed and compared, including the ratio of retention factor (*k*), the Tanimoto similarity, the difference of log D, the ratio of log P, and dual filters. Among the models, the filter that utilised the ratio of retention factors appeared to be the most effective approach to minimizing prediction errors. Therefore, in QSRR modelling the concept of chromatographic similarity should be considered and implemented using a measure of similarity that adequately reflects the retention of compounds. In other words, the design of other measures of similarity should select the same compounds in the dataset as does the *k*-ratio filter to establish accurate predictive QSRR models with acceptable prediction errors. Since the *k*-ratio filter is impractical, it cannot be used directly in QSRR modelling for retention prediction but is relevant as a benchmark for the minimum achievable prediction error in the QSRR modelling and to compare the performance of other practical filters.

The use of a Tanimoto filter based on the similarity of chemical structures resulted in much higher prediction errors when compared with the *k*-ratio similarity filter. Moreover, the low average Tanimoto score of the top ten compounds in the training set for each target compound in all three datasets also indicated the need for larger and homogenous datasets, allowing sufficient numbers of compounds with a high pair-wise similarity to be found when using the Tanimoto filter. Alternatively, the use of a Tanimoto cut-off score to identify the

best training set might be a better option. In general, the prediction errors were slightly improved using a log D filter which represents the hydrophobic interaction in reversed-phase retention, while the log P filter did not give the same improvement in error and therefore does not reflect hydrophobic interaction well. Similarly, the design of the dual filter, in which the Tanimoto filter or log D filter (the primary filter) was combined with a *k*-ratio filter (the secondary filter) did not lead to any improvements in retention prediction in all three datasets. Since the dual filter is still based on the similarity of chemical structures it could still be beneficial in the case of larger and more homogenous datasets by including compounds with greater chromatographic similarity.

3.5 References

1. Davis, J.M. and J.C. Giddings, *Statistical theory of component overlap in multicomponent chromatograms.* Analytical Chemistry, 1983. **55**(3): p. 418-424.
2. De Beer, M., F.d. Lynen, K. Chen, P. Ferguson, M. Hanna-Brown, and P. Sandra, *Stationary-phase optimized selectivity liquid chromatography: development of a linear gradient prediction algorithm.* Analytical Chemistry, 2010. **82**(5): p. 1733-1743.
3. Dong, M.W., *HPLC Column Standardization in Pharmaceutical Development: A Case Study.* LCGC Asia Pacific, 2016. **19**(3): p. 27-31.
4. Karmarkar, S., R. Garber, Y. Genchanok, S. George, X. Yang, and R. Hammond, *Quality by design (QbD) based development of a stability indicating HPLC method for drug and impurities.* Journal of Chromatographic Science, 2011. **49**(6): p. 439-446.
5. Talebi, M., S.H. Park, M. Taraji, Y. Wen, R.I. Amos, P.R. Haddad, R. Shellie, R. Szucs, C. Pohl, and J.W. Dolan, *Retention time prediction based on molecular structure in pharmaceutical method development: A perspective.* LCGC North America, 2016. **34**(8): p. 550-558.
6. Vogt, F.G. and A.S. Kord, *Development of quality-by-design analytical methods.* Journal of Pharmaceutical Sciences, 2011. **100**(3): p. 797-812.
7. Young, C.S. and R.J. Weigand, *An efficient approach to column selection in HPLC method development.* LCGC North America, 2002. **20**(5): p. 464-473.
8. Park, S.H., P.R. Haddad, M. Talebi, E. Tyteca, R.I. Amos, R. Szucs, J.W. Dolan, and C.A. Pohl, *Retention prediction of low molecular weight anions in ion chromatography based on quantitative structure-retention relationships applied to the linear solvent strength model.* Journal of Chromatography A, 2017. **1486**: p. 68-75.
9. Abate-Pella, D., D.M. Freund, Y. Ma, Y. Simón-Manso, J. Hollender, C.D. Broeckling, D.V. Huhman, O.V. Krokhin, D.R. Stoll, and A.D. Hegeman, *Retention projection enables accurate calculation of liquid chromatographic retention times across labs and methods.* Journal of Chromatography A, 2015. **1412**: p. 43-51.
10. Dong, M.W., *A Universal Reversed-Phase HPLC Method for Pharmaceutical Analysis.* LCGC North America, 2016. **34**(6): p. 408-419.
11. Hewitt, E.F., P. Lukulay, and S. Galushko, *Implementation of a rapid and automated high performance liquid chromatography method development strategy for pharmaceutical drug candidates.* Journal of Chromatography A, 2006. **1107**(1-2): p. 79-87.
12. García-Lavandeira, J., B. Losada, J. Martínez-Pontevedra, M. Lores, and R. Cela, *Computer-assisted method development in liquid chromatography–mass spectrometry: New proposals.* Journal of Chromatography A, 2008. **1208**(1): p. 116-125.
13. Zamora, I., F. Fontaine, B. Serra, and G. Plasencia, *High-throughput, computer assisted, specific MetID. A revolution for drug discovery.* Drug Discovery Today: Technologies, 2013. **10**(1): p. 199-205.

14. Bolanča, T., Š. Ukić, M. Novak, and M. Rogošić, *Computer assisted method development in liquid chromatography.* Croatica Chemica Acta, 2014. **87**(2): p. 111-122.
15. Tyteca, E., S.H. Park, R.A. Shellie, P.R. Haddad, and G. Desmet, *Computer-assisted multi-segment gradient optimization in ion chromatography.* Journal of Chromatography A, 2015. **1381**: p. 101-109.
16. Molnár, I., H.-J. Rieger, and R. Kormány, *Chromatography modelling in high performance liquid chromatography method development.* Chromatography Today, 2013. **6**(1): p. 3-8.
17. Kormány, R., J. Fekete, D. Guillarme, and S. Fekete, *Reliability of simulated robustness testing in fast liquid chromatography, using state-of-the-art column technology, instrumentation and modelling software.* Journal of Pharmaceutical and Biomedical Analysis, 2014. **89**: p. 67-75.
18. Tyteca, E., M. Talebi, R. Amos, S.H. Park, M. Taraji, Y. Wen, R. Szucs, C.A. Pohl, J.W. Dolan, and P.R. Haddad, *Towards a chromatographic similarity index to establish localized quantitative structure-retention models for retention prediction: use of retention factor ratio.* Journal of Chromatography A, 2017. **1486**: p. 50-58.
19. Schmidt, A.H., M. Stanic, and I. Molnár, *In silico robustness testing of a compendial HPLC purity method by using of a multidimensional design space build by chromatography modeling—Case study pramipexole.* Journal of Pharmaceutical and Biomedical Analysis, 2014. **91**: p. 97-107.
20. Borges, E.M., *How to select equivalent and complimentary reversed phase liquid chromatography columns from column characterization databases.* Analytica Chimica Acta, 2014. **807**: p. 143-152.
21. Claessens, H. and M. Van Straten, *Review on the chemical and thermal stability of stationary phases for reversed-phase liquid chromatography.* Journal of Chromatography A, 2004. **1060**(1): p. 23-41.
22. Taraji, M., P.R. Haddad, R.I. Amos, M. Talebi, R. Szucs, J.W. Dolan, and C.A. Pohl, *Prediction of retention in hydrophilic interaction liquid chromatography using solute molecular descriptors based on chemical structures.* Journal of Chromatography A, 2017. **1486**: p. 59-67.
23. Desfontaine, V., D. Guillarme, E. Francotte, and L. Nováková, *Supercritical fluid chromatography in pharmaceutical analysis.* Journal of Pharmaceutical and Biomedical Analysis, 2015. **113**: p. 56-71.
24. Kormány, R., I. Molnár, J. Fekete, D. Guillarme, and S. Fekete, *Robust UHPLC separation method development for multi-API product containing amlodipine and bisoprolol: the impact of column selection.* Chromatographia, 2014. **77**(17-18): p. 1119-1127.
25. Dolan, J., A. Maule, D. Bingley, L. Wrisley, C. Chan, M. Angod, C. Lunte, R. Krisko, J. Winston, and B. Homeier, *Choosing an equivalent replacement column for a reversed-phase liquid chromatographic assay procedure.* Journal of Chromatography A, 2004. **1057**(1): p. 59-74.
26. Euerby, M.R. and P. Petersson, *Chromatographic classification and comparison of commercially available reversed-phase liquid chromatographic columns containing polar embedded groups/amino endcappings using principal component analysis.* Journal of Chromatography A, 2005. **1088**(1): p. 1-15.
27. Kromidas, S., *HPLC made to measure: a practical handbook for optimization*. 2008: John Wiley & Sons.
28. Siouffi, A. and R. Phan-Tan-Luu, *Optimization methods in chromatography and capillary electrophoresis.* Journal of Chromatography A, 2000. **892**(1): p. 75-106.
29. Andrić, F. and K. Héberger, *How to compare separation selectivity of high-performance liquid chromatographic columns properly?* Journal of Chromatography A, 2017. **1488**: p. 45-56.

30. Ghasemi, J. and S. Saaidpour, *QSRR prediction of the chromatographic retention behavior of painkiller drugs.* Journal of Chromatographic Science, 2009. **47**(2): p. 156-163.
31. Tyrkkö, E., A. Pelander, and I. Ojanperä, *Prediction of liquid chromatographic retention for differentiation of structural isomers.* Analytica Chimica Acta, 2012. **720**: p. 142-148.
32. Kazakevich, Y.V. and R. Lobrutto, *HPLC for pharmaceutical scientists*. 2007: John Wiley & Sons.
33. Perisic-Janjic, N., R. Kaliszan, P. Wiczling, N. Milosevic, G. Uscumlic, and N. Banjac, *Reversed-phase TLC and HPLC retention data in correlation studies with in silico molecular descriptors and druglikeness properties of newly synthesized anticonvulsant succinimide derivatives.* Molecular Pharmaceutics, 2011. **8**(2): p. 555-563.
34. Leardi, R. and A.L. Gonzalez, *Genetic algorithms applied to feature selection in PLS regression: how and when to use them.* Chemometrics and Intelligent Laboratory Systems , 1998. **41**(2): p. 195-207.
35. Talebi, M., G. Schuster, R.A. Shellie, R. Szucs, and P.R. Haddad, *Performance comparison of partial least squares-related variable selection methods for quantitative structure retention relationships modelling of retention times in reversed-phase liquid chromatography*. Journal of Chromatography A, 2015. **1424**: p. 69-76.
36. Žuvela, P., J.J. Liu, K. Macur, and T. Baczek, *Molecular descriptor subset selection in theoretical peptide quantitative structure–retention relationship model development using nature-inspired optimization algorithms.* Analytical Chemistry, 2015. **87**(19): p. 9876-9883.
37. Gramatica, P., *Principles of QSAR models validation: internal and external.* Molecular Informatics, 2007. **26**(5): p. 694-701.
38. Muteki, K., J.E. Morgado, G.L. Reid, J. Wang, G. Xue, F.W. Riley, J.W. Harwood, D.T. Fortin, and I.J. Miller, *Quantitative structure retention relationship models in an analytical Quality by Design framework: simultaneously accounting for compound properties, mobile-phase conditions, and stationary-phase properties.* Industrial & Engineering Chemistry Research, 2013. **52**(35): p. 12269-12284.
39. Goryński, K., B. Bojko, A. Nowaczyk, A. Buciński, J. Pawliszyn, and R. Kaliszan, *Quantitative structure–retention relationships models for prediction of high performance liquid chromatography retention time of small molecules: endogenous metabolites and banned compounds.* Analytica Chimica Acta, 2013. **797**: p. 13-19.
40. *in, Talete srl, Dragon 6.0 for Windows (Software For Molecular Descriptor Calculations);* http://www.talete.mi.it/ Talete, Milano, Italy.
41. *R. Todeschini, V. Consonni, Handbook of Molecular Descriptors*. Wiley-WCH, Weinheim, 2000.
42. Tropsha, A., P. Gramatica, and V.K. Gombar, *The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models.* Molecular Informatics, 2003. **22**(1): p. 69-77.
43. Willett, P., *Similarity-based virtual screening using 2D fingerprints.* Drug Discovery Today, 2006. **11**(23-24): p. 1046-1053.
44. Park, S.H., P.R. Haddad, R.I. Amos, M. Talebi, R. Szucs, C.A. Pohl, and J.W. Dolan, *Towards a chromatographic similarity index to establish localised Quantitative Structure-Retention Relationships for retention prediction. III Combination of Tanimoto similarity index, logP, and retention factor ratio to identify optimal analyte training sets for ion chromatography*. Journal of Chromatography A, 2017. **1520**: p. 107-116.

4 Retention Prediction in Reversed-Phase Liquid Chromatography using Quantitative Structure-Retention Relationships: Application to the Hydrophobic Subtraction Model

4.1 Introduction

In many laboratory situations it is important to make an early assessment of the probable retention times of compounds. One significant example is in early stage drug discovery in the pharmaceutical industry when a variety of potential synthetic routes are being considered for a desired new active pharmaceutical compound. The evaluation of these synthetic routes must consider a range of parameters, including the likelihood of co-elution of predicted impurity compounds with each other or, more importantly, co-elution of these impurities with the active pharmaceutical compound. Failure to predict such co-elution may lead to costly or intractable analytical problems in downstream phases of drug development, especially in producing the required analytical data for registration of a new drug product. Computer tools that permit reasonably accurate prediction of retention times based only on the chemical structures of the compounds under consideration would clearly be highly beneficial in drug discovery applications, as well as in a wide range of other industries.

It is important to note that the goal of these predictions is not to identify the precise combination of mobile phase composition and stationary phase that would constitute the final chromatographic method for the separation of a group of target compounds. Rather, the objective is to derive rapidly, without experimentation, an overview of the retention behaviour of the target compounds on a range of stationary phases. It is recognised that a subsequent optimisation step, usually involving extensive experimentation, would be necessary before the final separation conditions are selected. Therefore, retention prediction tools used to obtain an overview of compound retention behaviour can be calculated from a retention database obtained under a very limited range of mobile phase conditions. In the present study, retention data obtained using a single mobile phase composition will be used.

As mentioned in the introduction, Quantitative Structure-Retention Relationships (QSRR) methodology aims to find meaningful relationships between chromatographic retention times and the chemical structure of compounds in the form of mathematical models [1]. A QSRR model is usually derived from a set of molecular descriptors, which are either determined experimentally (*e.g.* the log of the octanol-water partition coefficient (log P), molar refractivity, dipole moment, polarizability and other physico-chemical properties of compounds), or are computed theoretically from a symbolic representation of the molecules using molecular

modelling software, such as Dragon (Talete, Milano, Italy) or VolSurf+ (Molecular Discovery Ltd., Hertfordshire, UK).

There are two approaches to the prediction of retention using QSRR modelling. The first is based on a limited number of preselected descriptors which are calculated using built-in algorithms embedded into the software. ACD/ChromGenius (ACD/Labs, Toronto, Canada), for example, uses descriptors such as log P, the log of the compound distribution coefficient (log D), polar surface area, molecular volume, molecular weight, molar refractivity and the number of hydrogen bond donor and acceptor sites on the molecule as descriptors to build QSRR models. While the simple models generated using such tools are easy to interpret, this benefit could be offset (particularly in the case of more complex compounds) by a commensurate decrease in prediction accuracy arising from the simplification of the predicted retention mechanism due to the use of a very limited set of simple descriptors [2]. As an alternative, a large pool of molecular descriptors generated using molecular modelling software can be explored to extract the most relevant and informative descriptors. The appropriate incorporation of these additional descriptors has been shown in numerous studies to improve the prediction accuracy of models [3, 4]. Considering the large number of descriptors generated, implementing a suitable descriptor selection method, such as a genetic algorithm (GA), which is often combined with multiple linear regression (MLR) or partial least squares regression (PLS), becomes necessary to exclude noise from the model and reduce the risk of over-fitting and chance correlation [5, 6]. Multidimensional datasets often contain a much greater number of descriptors than compounds, and many of these descriptors are co-linear, redundant or can be regarded as noise [7, 8].

It has also been demonstrated in Chapter 3 and in other studies that QSRR prediction can benefit greatly from the localisation of predictive models, where individual models are built for each test compound using training sets formed from clusters of similar compounds in the dataset [8, 9]. Accordingly, different measures of similarity can be used, such as log D for physico-chemical similarity or Tanimoto similarity (TS) index for structural similarity based on 2D fingerprints [10]. While some improvements in prediction accuracy have become evident from these types of database filtering approaches, simple similarity measures like these appear to be less efficient for datasets containing compounds with more complex structures [3, 10]. This is because other compound properties (*e.g.*, charge state or hydrogen bonding) may also collectively contribute to the retention behaviour of compounds and need to be taken into account for more accurate filtering [3, 11].

A practical limitation that prevents widespread applicability of many QSRR models is the use of retention data for only a single stationary phase and mobile phase for modelling, which

consequently allows retention prediction for new compounds only within the confines of the specified conditions for the database. One solution to overcome this limitation might be to extend the scope of modelling by including more chromatographic conditions and columns in a structured experimental design. However, this approach is inevitably time and resource intensive. A more viable option with less demand on labour and time is modelling retention indirectly by predicting relevant coefficients in representative equations or models describing retention mechanisms, such as the Linear Solvent Strength (LSS) model in ion chromatography (IC) or the Hydrophobic Subtraction Model (HSM, see Eq. 4.1 below) in RPLC. In previous studies by the Haddad group, the benefits of this approach have been demonstrated in IC by predicting *a* and *b* coefficients in the LSS model and subsequently using these coefficients to predict the retention of ionised compounds over a wide range of eluent concentrations [12, 13]. In the development of the HSM, Snyder and co-workers concluded that retention and column selectivity in RPLC using alkyl silica phases (*i.e.*, C3-C18) are governed chiefly by five different physico-chemical properties, namely hydrophobicity, steric resistance, hydrogen-bond acidity, hydrogen-bond basicity, and ionic interactions between compound and stationary phase (see section 1.5.1 Chapter 1) [11, 14-18].

$$\log \alpha \equiv \log\left(\frac{k}{k_{\mathrm{EB}}}\right) = \eta'\mathbf{H} - \sigma'\mathbf{S}^* + \beta'\mathbf{A} + \alpha'\mathbf{B} + \kappa'\mathbf{C} \qquad 4.1$$

While the HSM was not originally aimed at retention prediction, a prediction accuracy of ±1-2% in *k* claimed using this model suggests that it has great potential for the purpose of retention prediction as well [15]. Additionally, such a high level of accuracy suggests that all major contributors to the retention mechanism in RPLC (using common C8 and C18 phases) are taken into account. Therefore, it makes sense to expect strong correlations between a solute's coefficients and its chemical properties on one hand, and similarly between column coefficients and column properties, such as ligand length (*i.e.*, C8, C18, *etc.*), particle pore diameter, *etc*. on the other [19]. At present, column coefficients for nearly 700 commercial C8 and C18 silica-based stationary phases characterised using the HSM are available through an open access database hosted by the U.S. Pharmacopeia Convention (USP) website [20, 21]. Therefore, there is a unique opportunity for predicting the retention of potentially any given compound on all of the characterised columns, albeit under only one isocratic condition, provided that the HSM solute coefficients are available for that compound. While experimental determination of solute coefficients is feasible through collecting the necessary retention data on at least five different columns, this approach is time and resource intensive. As an alternative, coefficients for any given compound might be predicted from their correlations with molecular descriptors derived only from the chemical structure of the compound.

The aim of this study was to demonstrate the utility of the proposed approach of modelling solute coefficients of the HSM, by fitting the predicted solute coefficients combined with the column coefficients into the HSM, allowing retention times of compounds to be predicted across a wide range of RPLC columns. The contribution of different resources of molecular descriptors was evaluated and compared. Additionally, the importance of the five terms of the HSM to retention prediction was investigated. Different approaches were used to cluster compounds into the training sets prior to deriving local QSRR models. The performance of each filtering approach in enhancing the prediction accuracy was then compared against the classical approach where a global QSRR model is derived using compounds in training set without filtering. The predictive power of the established models was evaluated using a series of criteria and statistical analysis and a proof of concept demonstration of the use of QSRR was performed by predicting the retention times of five representative compounds on nine columns for which HS coefficients were known.

4.2 Materials and methods

4.2.1 Datasets

Compounds from Dataset 1 [14, 15] for which five experimentally-derived solute coefficients (η', σ', β', α', κ') are reported, and compounds from Dataset 2 [22], for which only three solute coefficients (η', σ', β') were found sufficient for their representation [15, 22], were employed in the present study. In dataset 1 there are 63 neutral compounds, 13 acids and 12 bases, and in dataset 2 the compounds are all neutral. The Tanimoto similarity analysis indicates a large diversity of compounds in both datasets. Instead of modelling the retention of compounds in Dataset 1 for all ten columns, here, retention of compounds for the GL Inertsil ODS-3 column (Table 2.1, Chapter 2) was modelled.

Compounds in Dataset 2 were employed as well because a column was found in common between Datasets 1 and 2, which was a HP Zorbax SB C18 (Column 3, 150 mm × 4.6 mm i.d., 5µm) column (denoted as Zorbax StableBond C18 in Dataset 2) from Agilent Technologies (Newport, DE, USA). Due to the lack of information about the void time and the dimensions of the column used in Dataset 2, an estimated void time was calculated by correlating the retention factors of the overlapping compounds between these two databases and then using this void time to calculate retention times for the 60 remaining compounds in Dataset 2. Since small differences in solute coefficients were observed for the compounds in common, likely related to the 10 °C difference in column temperature employed during data collection for the two separate studies [15], *t*-tests were performed on solute coefficients of the compounds in common before combining the two datasets. Results confirmed the null hypothesis that the means of the two populations were indeed equal for all solute coefficients. Nevertheless, a

porting equation for each solute coefficient was derived by regressing the coefficient values in Dataset 1 against those in Dataset 2 and this equation was then used to readjust the values of solute coefficients for the remaining 60 compounds in Dataset 2 before merging them with Dataset 1 and obtaining the combined dataset. The coefficients of columns used in Dataset 1 (numbered from 1 to 10) and 2 (numbered from 11 to 15) are listed in Table 4.1. The solute coefficients of 88 compounds from Dataset 1 and 60 compounds from Dataset 2 (the overlapping compounds were subtracted) are shown in Table 4.2. The retention of these compounds on the corresponding columns has been described previously in the section 2.1.1, Chapter 2 (Tables 2.5 and 2.6).

Table 4.1. Coefficients of columns used in Dataset 1 and Dataset 2

Column	Column coefficients				
	H	S	A	B	C (pH 2.8)
C1. GL Inertsil ODS-3	1.0048	-0.0126	-0.1285	-0.0255	-0.3501
C2. Waters Symmetry C18	1.0498	-0.0588	0.0104	-0.0289	-0.2071
C3. HP Zorbax SB C18	0.9981	0.0211	0.2715	0.0064	0.0854
C4. HP Zorbax SB C18	0.9666	0.0418	0.2642	0.0093	0.0505
C5. HP Zorbax SB-300 C18	0.8945	0.0426	0.1092	0.0761	0.2204
C6. HP Eclipse XDB-C18	1.0355	-0.0084	-0.0202	-0.0325	0.0443
C7. YMC Pack Pro C18	1.0022	0.0022	-0.1362	-0.0128	-0.0960
C8. YMC Pack Pro C18	1.0195	-0.0077	-0.1317	-0.0105	0.0088
C9. YMC Pack Pro C18	1.0106	-0.0067	-0.1357	-0.0099	0.0135
C10. Supelco Discovery C18	0.9861	-0.0226	-0.1279	0.0163	0.1899
C11. Zorbax StableBond C18	0.9907	0.0118	0.3429		
C12. Zorbax Rx-C18	1.0651	-0.0557	0.3853		
C13. Hypersil C18	0.9635	-0.0065	0.0967		
C14. Hypersil C8	0.8536	0.0170	0.0409		
C15. Zorbax C8	0.8267	0.0055	0.0643		

Table 4.2. Solute coefficients of compounds used in Dataset 1 (Compound ID 1 to 88) and Dataset 2 (Compound ID 89 to 148)

ID	Name	Solute coefficients				
		η'	σ'	β'	α'	κ'
1	benzene	-0.42	-0.20	0.01	-0.04	-0.02
2	toluene	-0.21	-0.13	0.00	-0.01	-0.01
3	ethylbenzene	0.00	0.00	0.00	0.00	0.00
4	p-xylene	0.02	-0.12	0.00	0.01	0.00
5	propylbenzene	0.23	0.13	-0.01	0.03	0.00
6	naphthalene	-0.05	0.06	-0.02	0.15	-0.02
7	p-chlorotoluene	0.01	-0.09	0.01	0.15	-0.02
8	dichlorobenzene	0.02	-0.04	-0.02	0.14	-0.02
9	benzotrichloride	0.15	0.41	-0.05	0.13	-0.03
10	bromobenzene	-0.15	-0.05	-0.01	0.09	-0.03
11	1-nitropropane	-0.84	-0.04	0.01	-0.11	0.00
12	nitrobenzene	-0.58	0.32	-0.01	0.01	-0.04
13	p-nitrotoluene	-0.38	0.43	-0.01	0.04	-0.03
14	p-nitrobenzylchloride	-0.37	0.60	-0.03	0.02	-0.03
15	N-benzylformamide	-1.31	0.04	0.07	0.04	0.03
16	anisole	-0.47	0.04	0.00	-0.05	-0.02
17	benzylalcohol	-1.15	-0.14	0.01	-0.10	0.02
18	3-phenylpropanol	-0.87	0.01	0.02	0.12	0.02
19	5-phenylpentanol	-0.49	0.21	0.04	0.37	0.03
20	phenol	-1.03	-0.17	-0.02	-0.04	0.02
21	p-chlorophenol	-0.76	-0.04	-0.04	0.15	0.00
22	2,3-dihydroxynaphthalene	-0.93	-0.01	-0.11	0.61	-0.03
23	1,3-dihydroxynaphthalene	-1.04	-0.04	-0.06	0.20	0.00
24	eugenol	-0.55	0.12	-0.03	0.15	0.01
25	danthron	-0.02	0.47	-0.04	0.29	-0.04
26	n-propylformate	-0.87	-0.17	0.05	-0.19	0.01
27	methylbenzoate	-0.53	0.30	0.03	-0.04	-0.04
28	benzonitrile	-0.72	0.25	0.02	-0.02	-0.03
29	coumarin	-0.93	-0.55	-0.02	0.65	-0.04
30	acetophenone	-0.75	0.19	0.04	-0.05	-0.01
31	benzophenone	-0.18	0.66	-0.01	0.09	-0.03
32	cis-chalcone	-0.05	0.82	-0.02	0.07	-0.02
33	trans-chalcone	0.03	0.92	-0.03	0.18	-0.04
34	cis-4-nitrochalcone	-0.10	1.10	-0.04	0.07	-0.04
35	trans-4-nitrochalcone	0.02	1.43	-0.08	0.01	-0.04
36	cis-4-methoxychalcone	-0.10	0.97	-0.03	0.06	-0.03
37	trans-4-methoxychalcone	0.01	1.17	-0.06	0.13	-0.04
38	prednisone	-1.17	0.98	0.09	0.02	0.02
39	hydrocortisone	-1.15	0.97	0.05	0.10	0.03

40	mephenytoin	-0.96	0.11	-0.02	0.05	0.02
41	oxazepam	-0.86	0.02	-0.06	0.58	0.03
42	flunitrazepam	-0.63	0.75	-0.01	0.16	-0.02
43	5,5-dimethylhydantoin	-0.88	1.28	-0.05	-0.45	0.03
44	N,N-dimethylacetamide	-1.92	0.00	1.00	0.00	0.00
45	amitriptyline	-1.10	0.05	-0.03	0.32	0.83
46	diphenhydramine	-1.41	-0.06	0.00	0.16	1.02
47	D,L-propanolol	-1.65	-0.18	0.01	-0.33	1.23
48	nortriptyline	-1.17	0.06	-0.04	0.38	0.83
49	prolintane	-1.48	0.13	0.05	-0.53	1.08
50	4-n-pentylaniline	-0.50	-0.25	0.08	0.26	0.09
51	4-n-hexylaniline	-0.26	-0.21	0.08	0.42	0.09
52	4-n-heptylaniline	-0.02	-0.18	0.07	0.58	0.09
53	N-ethylaniline	-1.01	-0.41	0.06	-0.58	0.09
54	2-phenylpyridine	-0.69	0.21	0.05	-0.05	-0.01
55	diclofenac	-0.19	0.40	-0.04	0.86	-0.03
56	mefenamic	0.04	0.26	-0.04	0.92	-0.01
57	ketoprofen	-0.59	0.30	-0.04	0.55	0.01
58	diflunisal	-0.47	0.17	0.15	3.10	-0.43
59	4-n-butylbenzoic acid	-0.27	-0.28	0.02	1.02	0.04
60	4-n-pentylbenzoic acid	-0.05	-0.31	0.02	1.19	0.05
61	4-n-hexylbenzoic acid	0.18	-0.30	0.01	1.35	0.06
62	3-cyanobenzoic acid	-1.22	-0.06	0.03	0.91	-0.04
63	2-nitrobenzoic acid	-1.39	-0.19	0.02	1.45	-0.20
64	3-nitrobenzoic acid	-1.08	-0.02	0.05	1.21	-0.07
65	2,6-dimethylbenzoic acid	-0.93	-0.22	-0.02	0.46	0.01
66	2-fluorobenzoic acid	-1.15	-0.15	0.00	0.36	0.03
67	1,2-dinitrobenzene	-0.64	0.48	-0.03	0.12	-0.04
68	1,3-dinitrobenzene	-0.61	0.46	-0.03	0.02	-0.03
69	nitrocyclohexane	-0.39	0.31	0.00	-0.03	0.00
70	biphenyl	0.15	0.23	-0.04	0.10	-0.02
71	2-nitrobiphenyl	-0.11	0.80	-0.04	0.14	-0.04
72	3-nitrobiphenyl	0.08	0.75	-0.05	0.22	-0.04
73	2-biphenylmethanol	-0.51	0.37	-0.03	0.08	0.01
74	2,2'-biphenol	-0.69	0.15	-0.05	0.17	0.01
75	4,4'-biphenol	-1.08	0.49	-0.06	0.20	-0.01
76	diphenylbutyrolactone	-0.22	0.74	-0.02	0.09	-0.02
77	fluorescamine	-0.20	1.05	-0.04	0.15	-0.05
78	camphorquinone	-0.55	0.42	0.01	-0.09	0.01
79	N,N-diethylacetamide	-1.34	0.40	0.41	0.10	0.07
80	3-nitrophenol	-0.91	0.14	-0.04	0.16	-0.01
81	4-nitrophenol	-0.96	0.06	-0.03	0.22	-0.02
82	2,4-dinitrophenol	-0.82	-1.06	0.13	1.06	-0.16

83	2-5-dinitrophenol	-0.72	0.10	0.00	0.18	-0.04
84	picric acid	-0.93	-3.19	-0.09	4.04	-0.98
85	fisetin	-1.41	0.14	-0.07	0.41	0.01
86	biochanin A	-0.42	1.00	-0.09	0.24	-0.05
87	4-phenylpyridine	0.13	0.69	-0.05	0.17	-0.01
88	N-butylaniline	-0.33	0.24	-0.03	-0.36	0.02
89	1-butanol	-1.28	-0.47	0.17		
90	1-hexanol	-0.79	-0.06	0.12		
91	1-octanol	-0.32	0.34	0.09		
92	isopropanol	-1.63	-0.95	0.22		
93	cyclohexanol	-1.13	-0.23	0.22		
94	1-butanal	-0.98	-0.11	0.23		
95	1-hexanal	-0.50	0.36	0.16		
96	1-heptanal	-0.27	0.49	0.15		
97	1-octanal	-0.05	0.49	0.19		
98	N,N-dimethyl formamide	-1.98	-0.22	0.84		
99	N,N-diethyl formamide	-1.47	0.23	0.46		
100	N,N-dibutyl formamide	-0.54	1.05	0.18		
101	n-butyl acetate	-0.54	0.17	0.06		
102	n-amyl acetate	-0.31	0.36	0.03		
103	n-hexyl acetate	-0.08	0.50	0.03		
104	ethyl propionate	-0.77	-0.04	0.06		
105	ethyl butyrate	-0.53	0.09	0.03		
106	ethyl ether	-0.96	-0.56	0.22		
107	di-n-propyl ether	-0.35	-0.26	0.11		
108	di-n-butyl ether	0.16	0.10	0.08		
109	dioxane	-1.60	-0.54	0.49		
110	acetone	-1.46	-0.61	0.20		
111	butane-2-one	-1.18	-0.37	0.13		
112	heptane-2-one	-0.47	0.28	0.04		
113	nonane-2-one	-0.01	0.55	0.04		
114	cyclopentanone	-1.18	-0.11	0.22		
115	n-propionitrile	-1.20	-0.26	0.05		
116	n-valeronitrile	-0.75	0.19	-0.01		
117	n-hexanitrile	-0.52	0.33	-0.03		
118	n-hexyl cyanide	-0.26	0.49	-0.18		
119	n-heptyl cyanide	-0.07	0.62	-0.09		
120	n-octyl cyanide	0.16	0.71	-0.10		
121	n-nitrobutane	-0.58	0.25	-0.06		
122	n-nitropentane	-0.35	0.41	-0.08		
123	methylene chloride	-0.73	-0.17	-0.03		
124	chloroform	-0.48	-0.02	-0.04		
125	dibromomethane	-0.61	-0.08	0.00		

126	2-phenyl ethanol	-1.04	-0.14	0.04
127	3-phenyl propanol	-0.88	0.01	0.04
128	benzaldehyde	-0.77	0.14	0.07
129	ethyl benzoate	-0.31	0.32	-0.01
130	propiophenone	-0.50	0.25	0.06
131	m-toluenitrile	-0.50	0.44	0.01
132	benzyl cyanide	-0.68	0.33	-0.04
133	m-nitrotoluene	-0.34	0.46	-0.04
134	o-nitrotoluene	-0.39	0.45	-0.05
135	p-nitrobenzyl bromide	-0.30	0.69	-0.05
136	fluorobenzene	-0.41	-0.04	-0.04
137	chlorobenzene	-0.21	-0.09	-0.02
138	lodobenzene	-0.06	-0.02	0.02
139	benzyl bromide	-0.21	0.40	-0.05
140	p-bromotoluene	0.06	-0.06	0.01
141	n-butylbenzene	0.45	0.35	-0.02
142	tert.-butylbenzene	0.31	0.43	-0.03
143	mesitylene	0.21	0.02	0.02
144	anthracene	0.33	-0.35	0.07
145	m-cresol	-0.83	0.01	-0.06
146	p-cresol	-0.83	0.01	-0.06
147	o-cresol	-0.77	0.05	-0.07
148	p-ethylphenol	-0.64	0.10	-0.07

4.2.2 Calculation of the molecular descriptors

In the present study, different resources of molecular descriptors were employed and compared to improve the performance of the QSRR models. Molecular descriptors were calculated either using Dragon 6.0 (Talete, Milano, Italy) or VolSurf+ 1.0.7.1 (Molecular Discovery Ltd., Hertfordshire, UK), as detailed in section 2.3.2, Chapter 2. Besides using molecular descriptors generated from Dragon and VolSurf+ separately, a combination of molecular descriptors from both was also generated and tested.

4.2.3 Filtering approaches for Dataset 1

In this study, both global and local models (as described in section 2.3.5, Chapter 2) were created and compared for the prediction of solute coefficients. The five solute coefficients of the HSM were predicted as functions of molecular descriptors for their subsequent use in retention prediction. A popular version of a GA-PLS algorithm, originally written by Leardi [23] in Matlab (The Mathworks Inc., Natick, MA, USA) was modified and used for descriptor selection and modelling solute coefficients. To cope with the variability of the results arising from the intrinsic random selection nature of the genetic algorithm, the GA-PLS modelling

was repeated five times and the results were averaged. This part of work was performed for the 88 compounds in Dataset 1 for the GL Inertsil ODS-3 column with different resources of molecular descriptors (Dragon, VolSurf+ and the combined descriptors).

Leave-one-out (LOO) filtering: In this filtering method one compound was taken out as the target compound, the rest of the compounds were used as a training set for QSRR modelling. In this way, each target compound has its own local model for the prediction of five solute coefficients, created using information from the compounds in the entire dataset. Different resources of molecular descriptors were investigated and compared.

The Global approach: rather than building a model for each compound, another option for QSRR modelling which is easy to interpret is building just one model for all the compounds using a randomly allocated training set while the rest of the compounds in the database were treated as an external test set. In this work, a D-optimal algorithm suggested by Todeschini *et al*. [24] was used to allocate compounds into a training set (70%) and a test set (30%), respectively. This distance-based selection approach ensures homogenous sampling from a database leading to a uniform distribution of compounds between the resulting subsets [24]. Compounds in the training set were used to build QSRR models through GA-PLS, and compounds in the test set were employed to evaluate the predictive ability of the constructed QSRR models. Again, different resources of molecular descriptors were investigated and compared.

Local Compound Type (LCT) filtering: Another approach for QSRR modelling involves the classification of compounds. In this way, instead of building a local model for each compound separately, or one model to describe the whole dataset, a model was built for a group of compounds which lay within the same classification. For LCT, compounds were clustered according to their type (bases, acids, and neutrals), therefore, compounds belonging to the same type were classified into the same cluster. Each cluster was then divided, as for the global model, into a training set and a test set. A QSRR model was then derived for each cluster for the prediction of solute coefficients and the resources of molecular descriptors compared.

Local Second Dominant Interaction (LSDI) filtering: in Wilson's work [25], the HSM solute coefficients were experimentally generated based on the classification of compounds according to the interaction between compounds and stationary phase. "Ideal compounds" for which the retention was determined entirely by the hydrophobic interaction were used to yield the hydrophobicity coefficient [25]. Then, other solute coefficients were generated using a group of compounds that have been clustered according to their secondary dominant interaction after the hydrophobic interaction. Thus, compounds in Wilson's study were

allocated to different clusters, each corresponding to one of the terms in the HSM [25]. Following the subtraction of the effect of hydrophobicity, retention for compounds in each cluster was assumed to be predominantly influenced by the type of interaction linked to that cluster. Five clusters were identified, namely η' (hydrophobicity only), σ' (steric bulk), β' (hydrogen bonding basicity), α' (hydrogen bonding acidity), and κ' (charged compound) clusters, containing 25, 21, 4, 16 and 7 compounds, respectively [25]. Compounds for which retention appeared to be substantially influenced by more than one type of interaction (excluding hydrophobicity) were not assigned to any cluster by Wilson and co-workers, but have been allocated into a separate cluster (cluster 6) in the present study. After compound classification using this approach, and separation of each cluster into training and test sets, a QSRR model was built for each cluster and the solute coefficients were predicted and the resources of molecular descriptors compared.

4.2.4 Filtering approaches for the combined dataset

The prediction of ported solute coefficients and retention times was also performed on a combined dataset which contained 148 compounds (88 compounds from Dataset 1 and 60 compounds from Dataset 2). Again, both global and local QSRR models were built through PLS equations using a Matlab platform. This part of the study was performed for the retention data of the combined 148 compounds on a common column with molecular descriptors generated using VolSurf+ only.

Local Tanimoto Similarity (LTS) and Local Log D (LLD) filtering: previous results have shown that with a group of compounds in a training set having a sufficient level of similarity to the target, acceptable performance of retention prediction can be obtained [26]. In the LTS method, filtering was performed based on the Tanimoto Similarity (TS) index where the dataset was sorted based on the compounds' pairwise TS indices in relation to the target. Then, the top five compounds with pairwise TS indices of at least 0.5 were used as a training set to derive a separate QSRR model [26, 27]. If five compounds with a TS index of greater than 0.5 could not be found, the compound was not modelled.

In the LLD, compounds were sorted based on the ratio of their log D value to the log D of the target. Then, the top five compounds with log D ratio less than 1.1 were used as the training set [26]. Other conditions were the same as for the LTS approach. It is worth pointing out that, using either the LTS or the LLD approach, each compound has its own separate model for the prediction of solute coefficients.

The Global approach: this approach used all the compounds in the dataset to build one global model without any compound classification for the retention prediction of all

compounds [27]. Here, global PLS models (using 126 or 34 VolSurf+ descriptors, refer to the G126 and G34 below) were derived as the benchmark to gauge the improvements that the implementation of GA-PLS could bring to the final models. Before the modelling process, a D-optimal approach was used for splitting the combined dataset into a training set and an external test set. Then, a global QSRR model was built to predict solute coefficients and retention by fitting the predicted solute coefficients and their complementary column coefficients into the HSM.

Local Compound Type (LCT) filtering: the LCT filter as mentioned above was also applied to the 148 combined compounds based on their chemical nature, three clusters containing 13 acids, 12 bases, and 123 neutrals were obtained, respectively. Approximately 70% compounds of each cluster were then selected using a D-optimal approach for modelling and the remaining compounds were used for the external validation of the corresponding model.

Local Second Dominant Interaction (LSDI) filtering: as previously described, six clusters were obtained for 88 compounds from Dataset 1 using this filtering approach. For modelling the combined 148 compounds, the β' cluster, representing compound hydrogen-bond basicity, was expanded to seven by adding three more compounds with the same property from Dataset 2. Similarly, 70% of compounds in each cluster were allocated into the training set using a D-optimal algorithm, where the rest of the compounds in each cluster were taken as test sets. The applicability of this approach for predicting retention time for new test compounds was also demonstrated by using Dataset 2 as an external test set. To allocate new test compounds into the corresponding cluster so that the correct model could be used, TS searching was introduced. The structural similarity of each compound in Dataset 2 was investigated against training compounds (Dataset 1) in each cluster with the aim of finding one training compound with a pairwise Tanimoto structural similarity of at least 0.5. If such a similar compound was found the target compound was assigned to the same LDSI cluster as the compound with the greatest pairwise similarity. In total, 28 compounds out of 57 in Dataset 2 were assigned to clusters as follows: η' cluster (19 compounds), σ' cluster (6 compounds) and cluster 6 (3 compounds). The other 29 compounds in Dataset 2 were excluded since their pairwise TS indices were less than 0.5 when calculated against every compound in Dataset 1.

4.2.5 Statistics

The coefficient of determination (R^2), the slope of the regression with no forced intercept and root-mean-squared error of prediction (RMSEP) were used to evaluate model fitness with the requirement for the slope to be within the range of 0.85 to 1.15 [28]. The percentage root-mean-square error of prediction (RMSEP%) of retention time for the test set

was measured to externally validate the accuracy of GA-PLS models generated from the training set. The equations of the RMSEP and the RMSEP% were detailed in section 2.3.6, Chapter 2 (Eq. 2.2 and Eq. 2.3).

The predictive ability of the models was evaluated by inspecting the Regression Error Characteristic (REC) curves obtained by plotting the prediction error range against the percentage of data points predicted within that range [29, 30]. The null model, which can be regarded as the baseline model, was obtained by using the mean of the dependent variable (response) as a naïve predicted value for all compounds. REC curves were used to show the differences between regression models and have the advantage that the ranking of models is independent of the error measure used [30].

In addition, the overall performance of all generated models was further compared using the sum of ranking difference (SRD) approach [31-33] where parameters for each model were compared to a series of reference values, and each model ranked according to how large was the difference between its parameters and the reference values. The rankings were also compared to a confidence interval generated by using randomly ranked numbers [32, 34]. This ordering method provides a simple way to evaluate the models by comparing their SRD values (the closer the SRD value to zero, the better the approach) [31].

4.3 Results and Discussion

4.3.1 QSRR Prediction for Dataset 1

For the LOO approach, 88 compounds in Dataset 1 were used for the prediction of the five solute coefficients through QSRR models with different resources of molecular descriptors. The combined molecular descriptors were performed by running the GA to choose the Dragon descriptors first, and then running the same process to select VolSurf+ descriptors. Finally, the selected descriptors from Dragon and VolSurf+ were combined as a pool of combined descriptors for modelling. Figures 4.1 – 4.5 show the performance of solute coefficient predictions using three resources of descriptors.

Among the prediction of five solute coefficients, only the predictions of η' were acceptable based on the generated slope and correlation coefficient. Data points of prediction for the other four solute coefficients were either scattered on both sides of the trend line, or aggregated around the origin, and showed poor correlation between the measured and predicted solute coefficients. For η' prediction, as can be seen from Figure 4.1, a range of correlation coefficients between 0.6388 to 0.6894 was obtained with RMSEP values ranging from 0.28 to 0.30. Very similar results (the slope, the R^2, the RMSEP) were observed for all the different resources of descriptors using the same LOO approach, suggesting that there was

no significant improvement gained from using diverse types of descriptors or even from the combined descriptors. The same trend was also observed from the prediction of the other four solute coefficients, with almost identical plots being generated between measured and predicted coefficients for each resource of molecular descriptors. Additionally, regardless of the resource of descriptors used, or which solute coefficients were predicted, the same compounds were found to be outliers – *i.e.*, poorly predicted with high prediction errors.

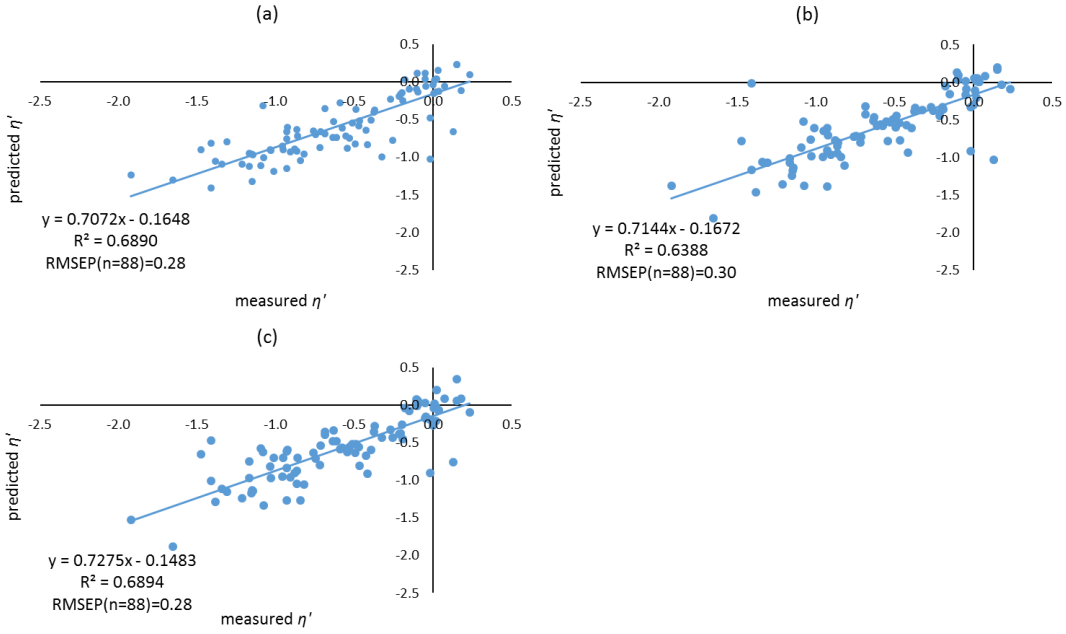


Figure 4.1. Solute coefficient (η') prediction using the LOO approach with (a) Dragon descriptors, (b) VolSurf+ descriptors, and (c) combined descriptors.

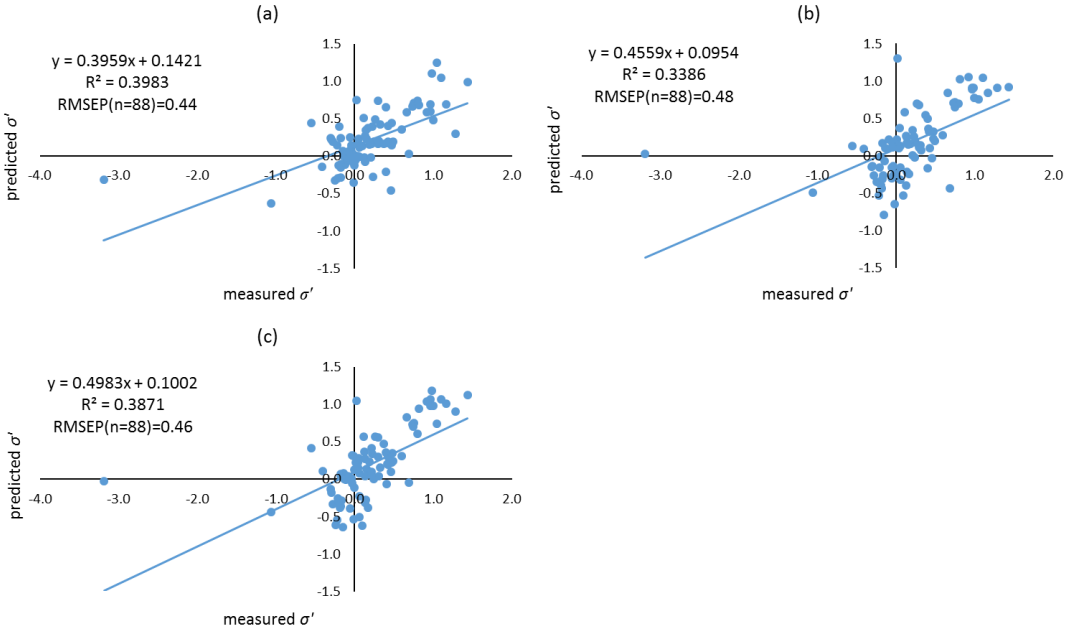


Figure 4.2. Solute coefficient (σ') prediction using the LOO approach with (a) Dragon descriptors, (b) VolSurf+ descriptors, and (c) combined descriptors.

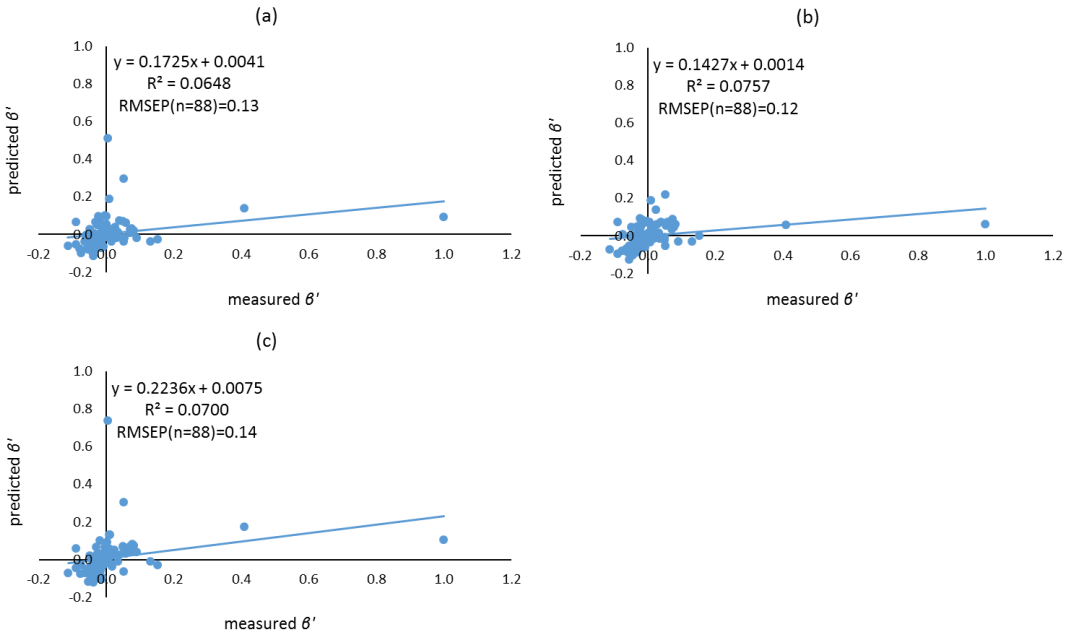


Figure 4.3. Solute coefficient (β') prediction using the LOO approach with (a) Dragon descriptors, (b) VolSurf+ descriptors, and (c) combined descriptors.

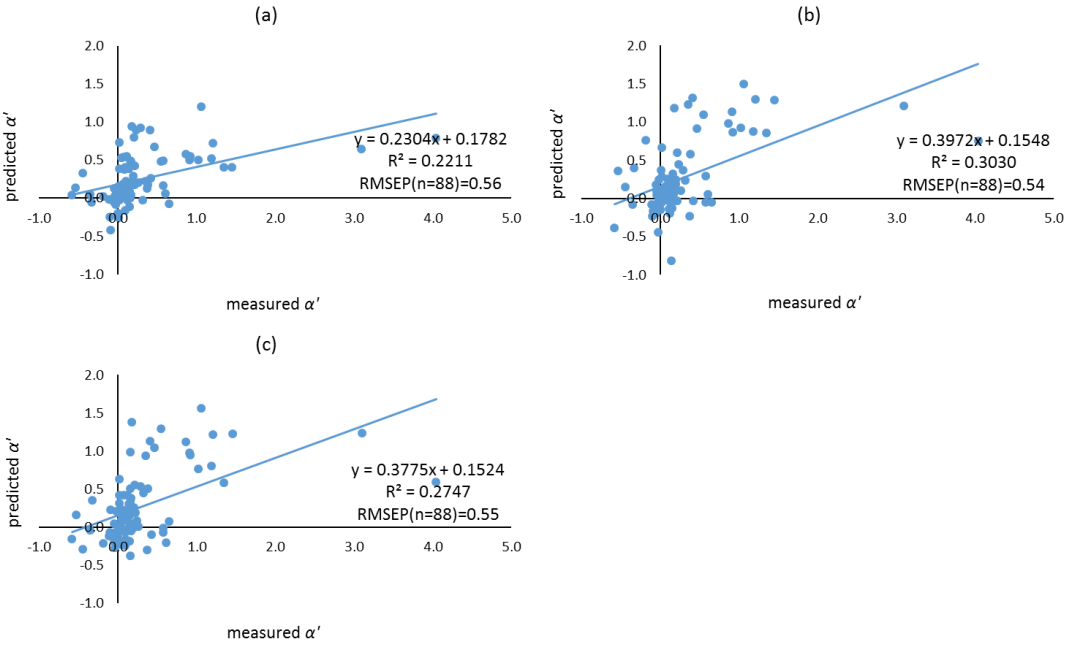


Figure 4.4. Solute coefficient (α') prediction using the LOO approach with (a) Dragon descriptors, (b) VolSurf+ descriptors, and (c) combined descriptors.

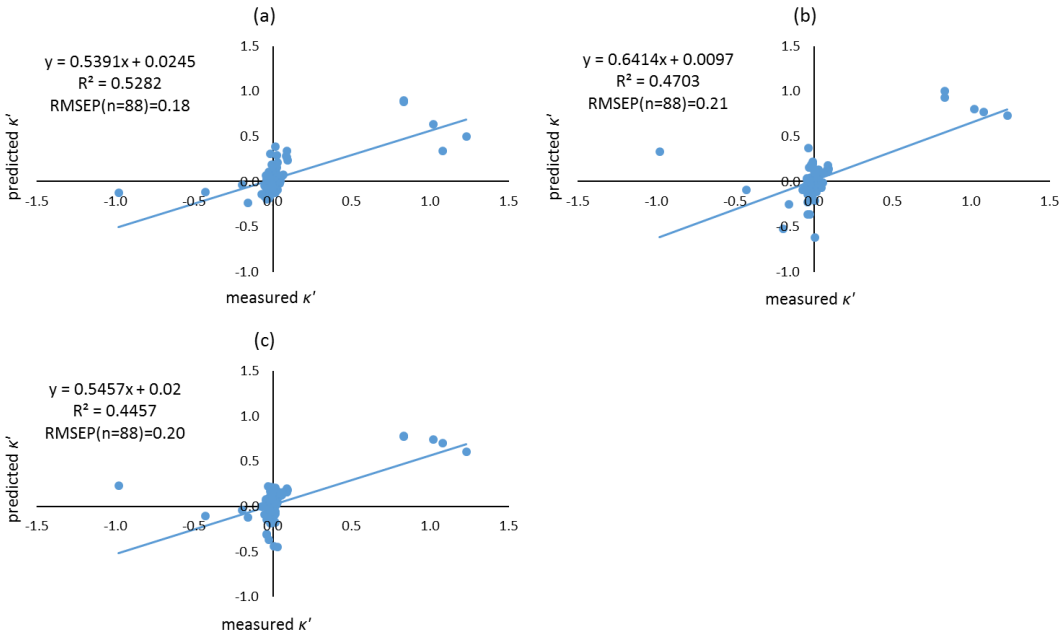


Figure 4.5. Solute coefficient (κ') prediction using the LOO approach with (a) Dragon descriptors, (b) VolSurf+ descriptors, and (c) combined descriptors.

Unlike the LOO approach, where each target compound has its own model, a group of compounds can be predicted using just one global model where training and test sets are employed. As shown in Figures 4.6 – 4.10, five solute coefficients were predicted, and the RMSEP errors are reported. Compared to the LOO approach, a slight improvement of prediction error was observed for the prediction of η' coefficient, RMSEP values of 0.28, 0.20, and 0.23 were obtained using Dragon, VolSurf+, and combined descriptors, respectively, while using the LOO approach but the same resources of descriptors resulted in RMSEP of 0.28, 0.30, and 0.28 for η' coefficient prediction. This indicated again that the resource of molecular descriptors has almost no influence on the prediction of solute coefficients. In terms of the prediction of the other four coefficients, as in the performance of the LOO approach, poorly correlated coefficients and much higher prediction errors were obtained with the global model.

As can be seen from the above approaches, neither the LOO nor the Global approach yielded acceptable prediction accuracy of solute coefficients. One thing in common between these two approaches is that compounds used for QSRR model construction were quite diverse with a wide range of molecular characteristics. In the LOO approach, one compound was taken out as the target analyte while the rest were used as the training set. In terms of the Global approach, 70% compounds of the whole dataset were used as the training set. Since this dataset is highly diverse (based on the Tanimoto similarity analysis), in the QSRR models generated using the LOO or the Global approach, compounds with widely diverse molecular characteristics were included, causing much higher prediction errors.

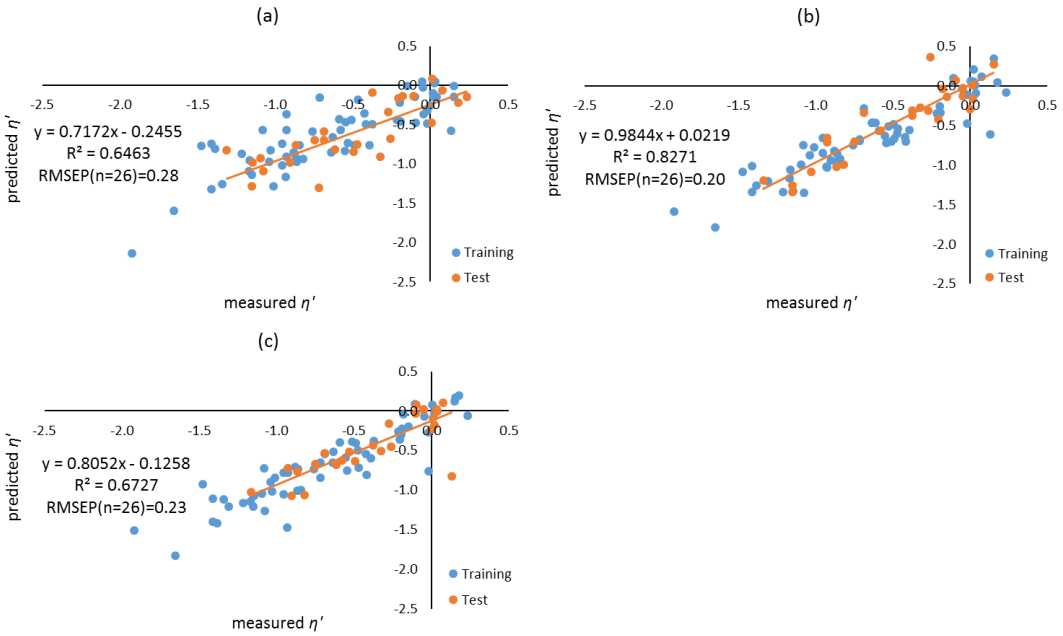


Figure 4.6. Solute coefficient (η') prediction using the Global approach with (a) Dragon descriptors, (b) VolSurf+ descriptors, and (c) combined descriptors.

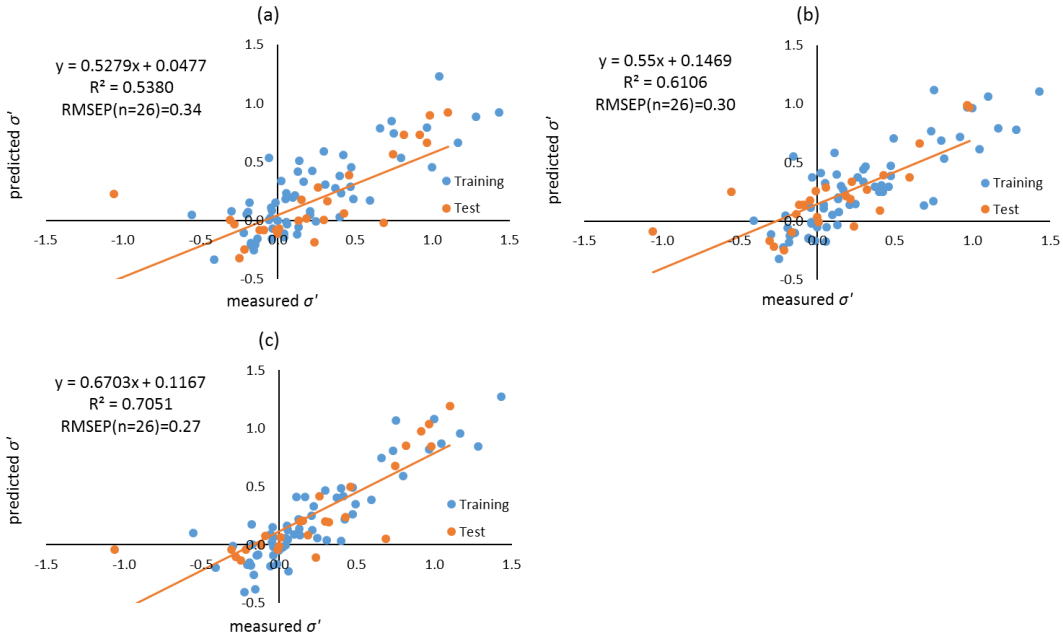


Figure 4.7. Solute coefficient (σ') prediction using the Global approach with (a) Dragon descriptors, (b) VolSurf+ descriptors, and (c) combined descriptors.

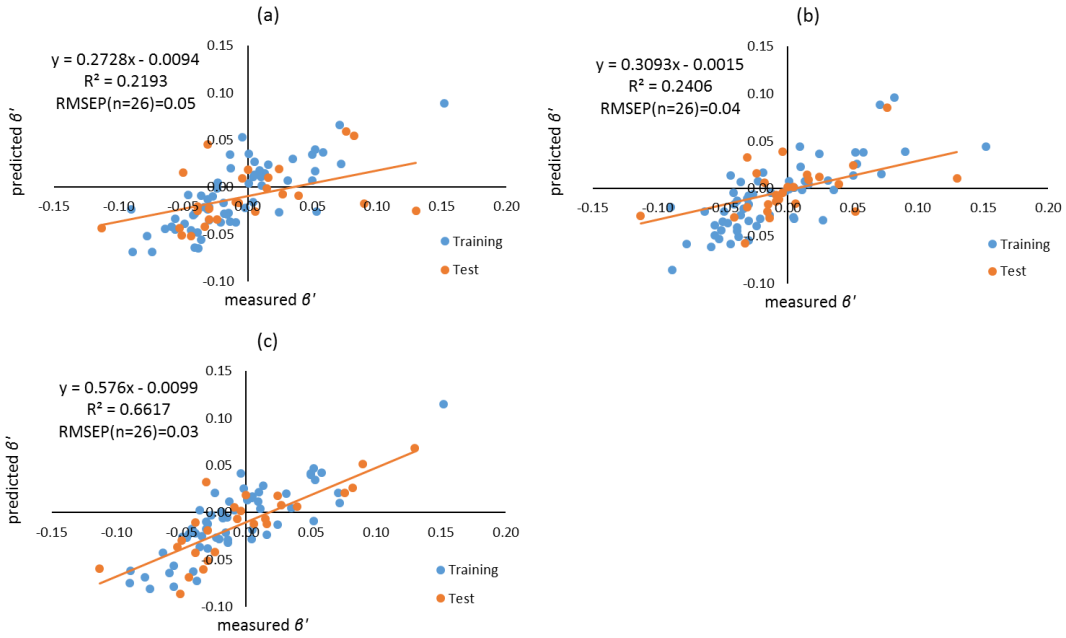


Figure 4.8. Solute coefficient (β') prediction using the Global approach with (a) Dragon descriptors, (b) VolSurf+ descriptors, and (c) combined descriptors.

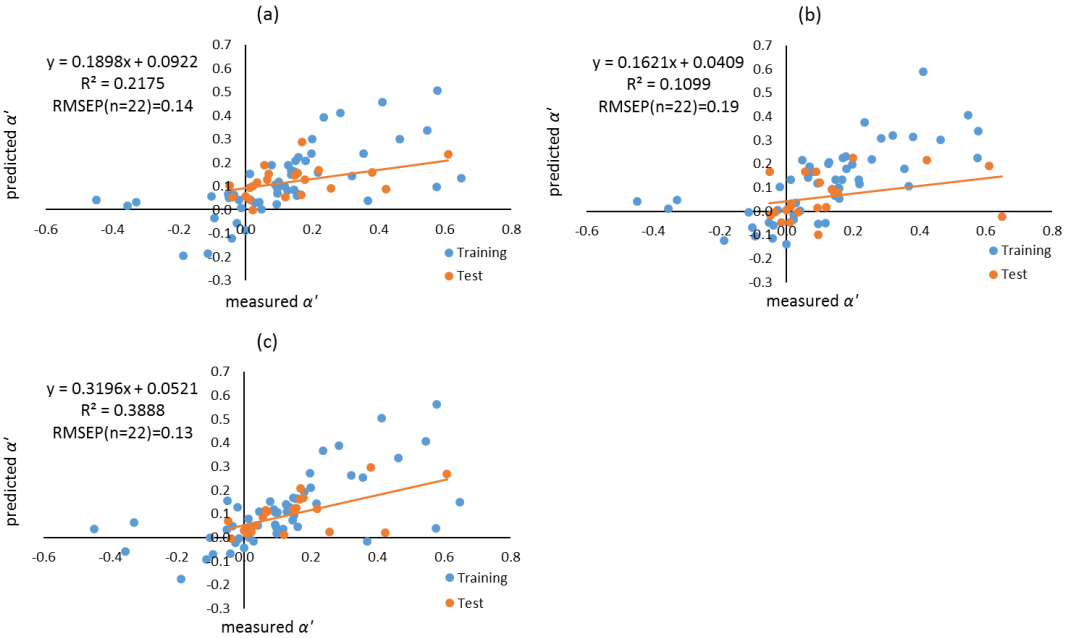


Figure 4.9. Solute coefficient (α') prediction using the Global approach with (a) Dragon descriptors, (b) VolSurf+ descriptors, and (c) combined descriptors.

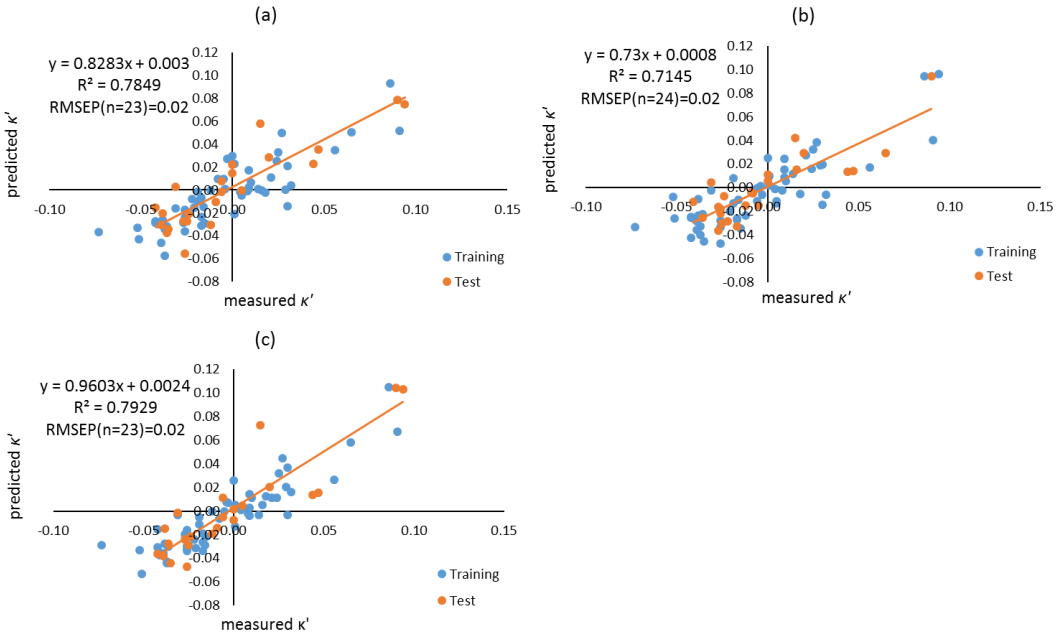


Figure 4.10. Solute coefficient (κ') prediction using the Global approach with (a) Dragon descriptors, (b) VolSurf+ descriptors, and (c) combined descriptors.

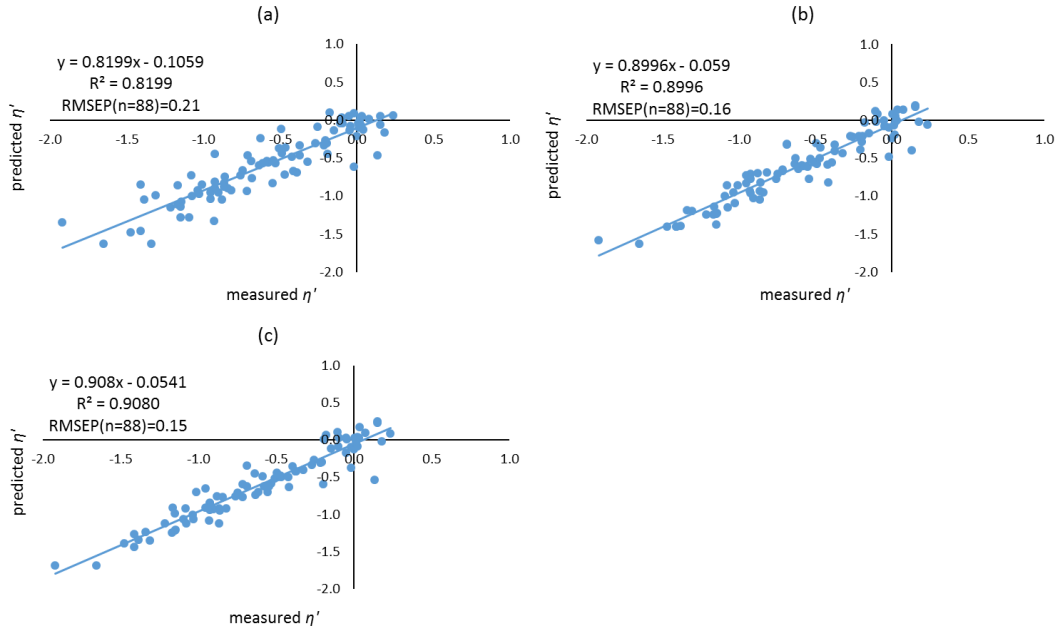


Figure 4.11. Solute coefficient (η') prediction using the LCT approach with (a) Dragon descriptors, (b) VolSurf+ descriptors, and (c) combined descriptors.

It has been shown that smaller, more similar training sets to the target compounds leads to greater prediction accuracy. Compound classification based on the type of the molecules was therefore investigated. The predictions for the five solute coefficients are listed in Figures 4.11 – 4.15. The accuracy of prediction for all the five solute coefficients was improved significantly compared to the LOO and the Global approach. In terms of the comparison of molecular descriptors, Dragon, VolSurf+, and combined descriptors demonstrated comparable results in the prediction of the five solute coefficients.

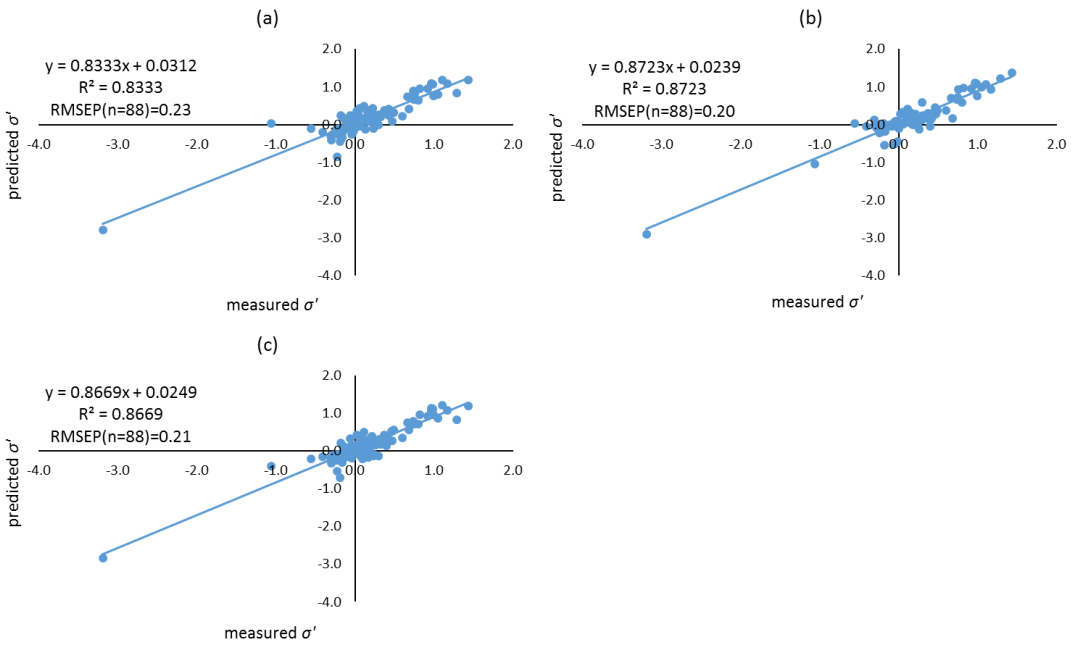


Figure 4.12. Solute coefficient (σ') prediction using the LCT approach with (a) Dragon descriptors, (b) VolSurf+ descriptors, and (c) combined descriptors.

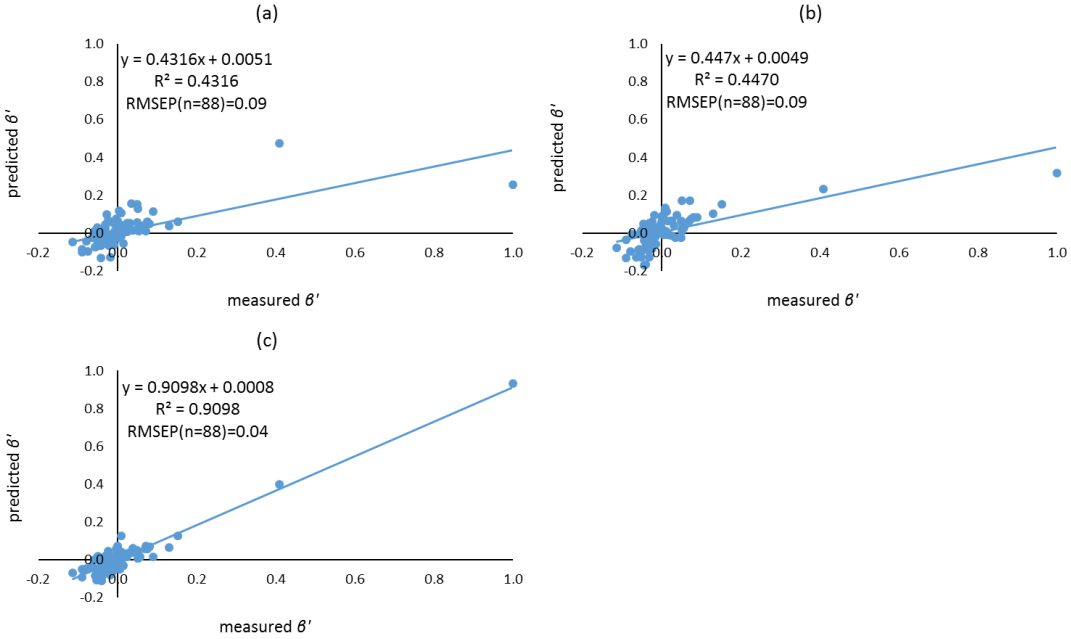


Figure 4.13. Solute coefficient (β') prediction using the LCT approach with (a) Dragon descriptors, (b) VolSurf+ descriptors, and (c) combined descriptors.

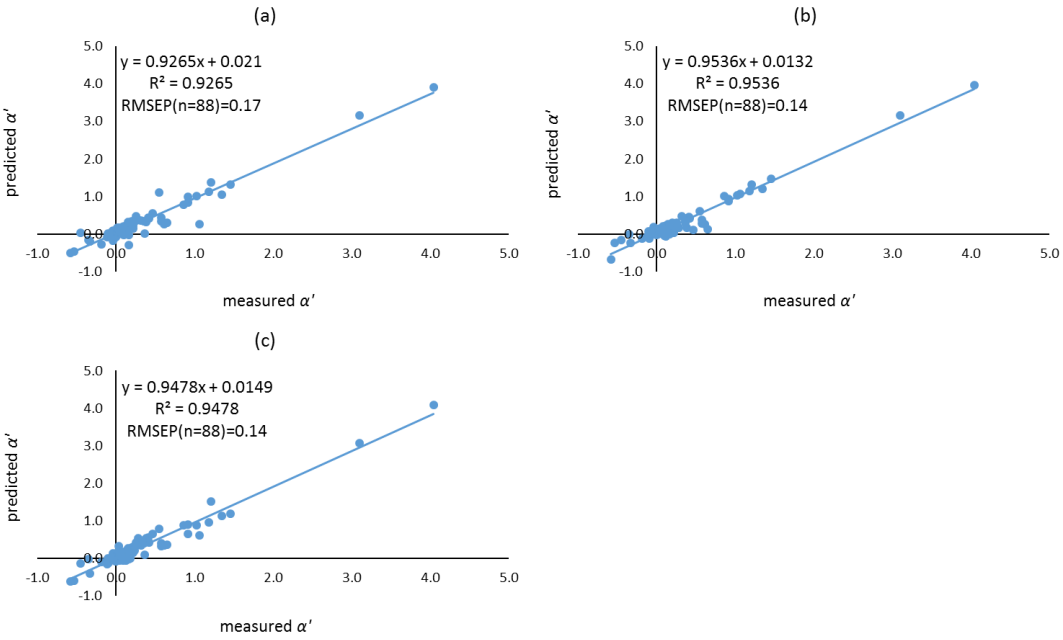


Figure 4.14. Solute coefficient (α') prediction using the LCT approach with (a) Dragon descriptors, (b) VolSurf+ descriptors, and (c) combined descriptors.

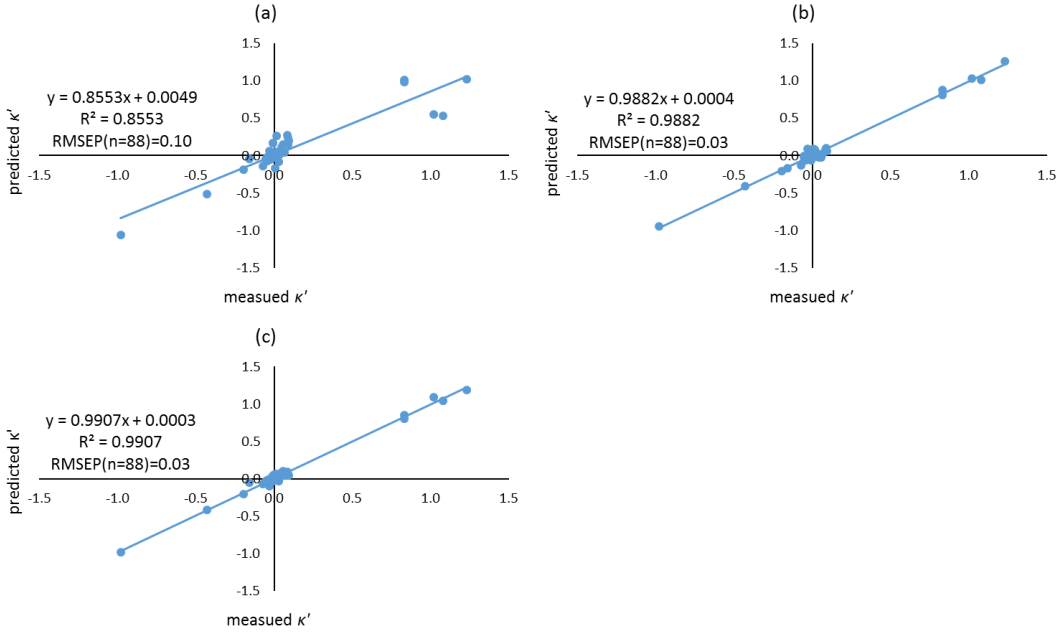


Figure 4.15. Solute coefficient (κ') prediction using the LCT approach with (a) Dragon descriptors, (b) VolSurf+ descriptors, and (c) combined descriptors.

Results obtained for the LCT approach suggested that better selection of the training set might improve the overall prediction accuracy of the models. In this context, the LSDI approach was investigated. Prediction of the five solute coefficients was performed using the same GA-PLS method and the results can be seen in Figures 4.16 – 4.20. Compared to the LCT approach, the LSDI approach improved the performance of the QSRR models. Linear correlations between measured and predicted solute coefficients were obtained and supported

by the distribution of data points around the trend line (much higher R^2 values were obtained). Again, a change in the resource of molecular descriptors made no difference to the prediction of coefficients, which consistent with the previous results (the LOO method, the Global method, and the LCT method).

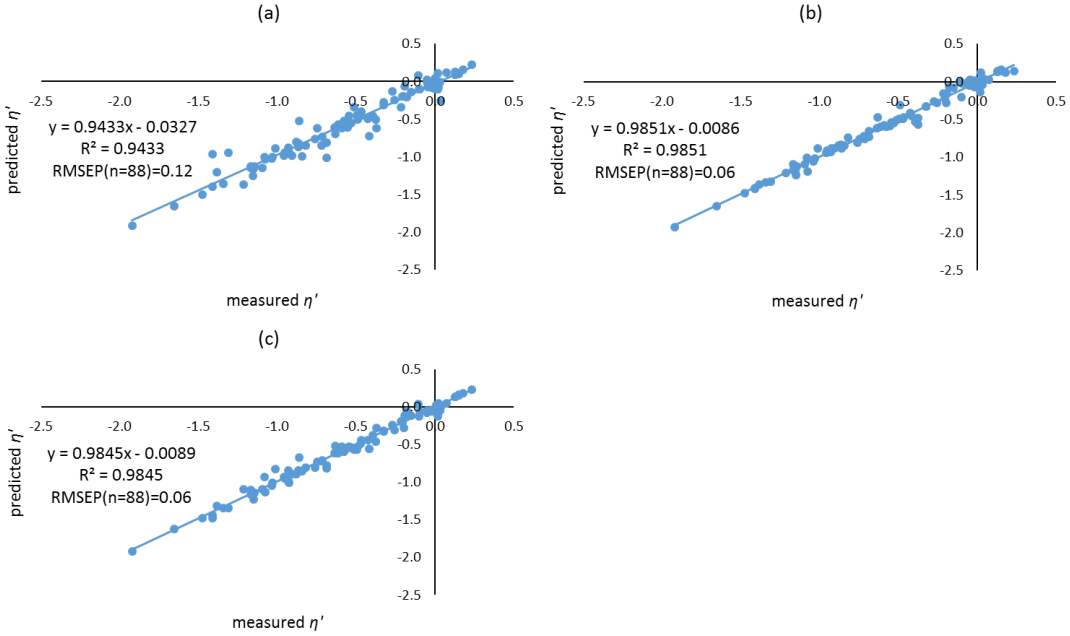


Figure 4.16. Solute coefficient (η') prediction using the LSDI approach with (a) Dragon descriptors, (b) VolSurf+ descriptors, and (c) combined descriptors.

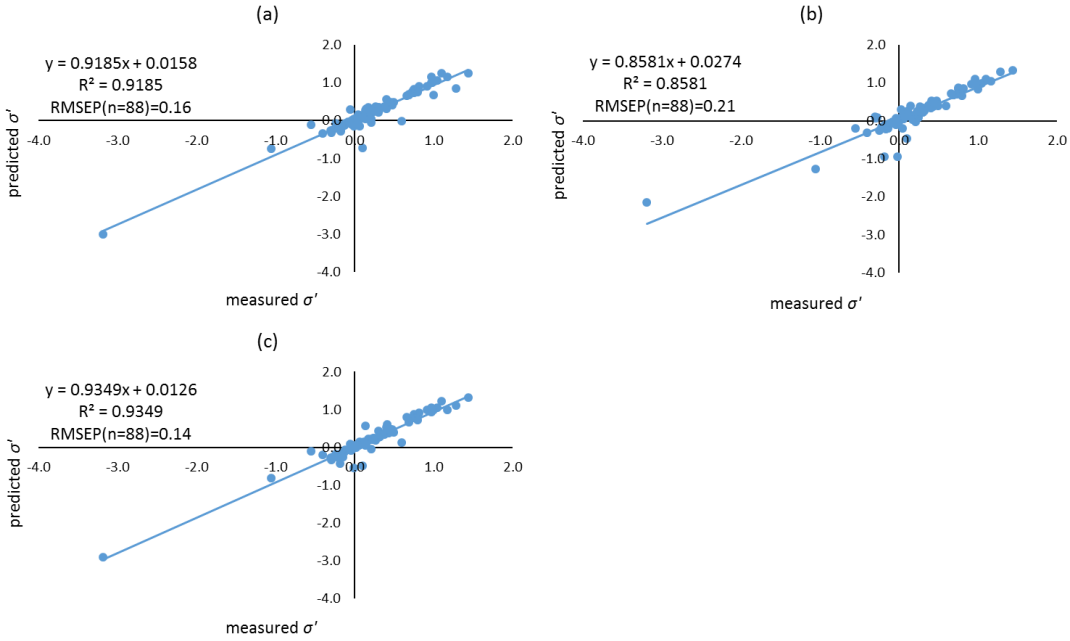


Figure 4.17. Solute coefficient (σ') prediction using the LSDI approach with (a) Dragon descriptors, (b) VolSurf+ descriptors, and (c) combined descriptors.

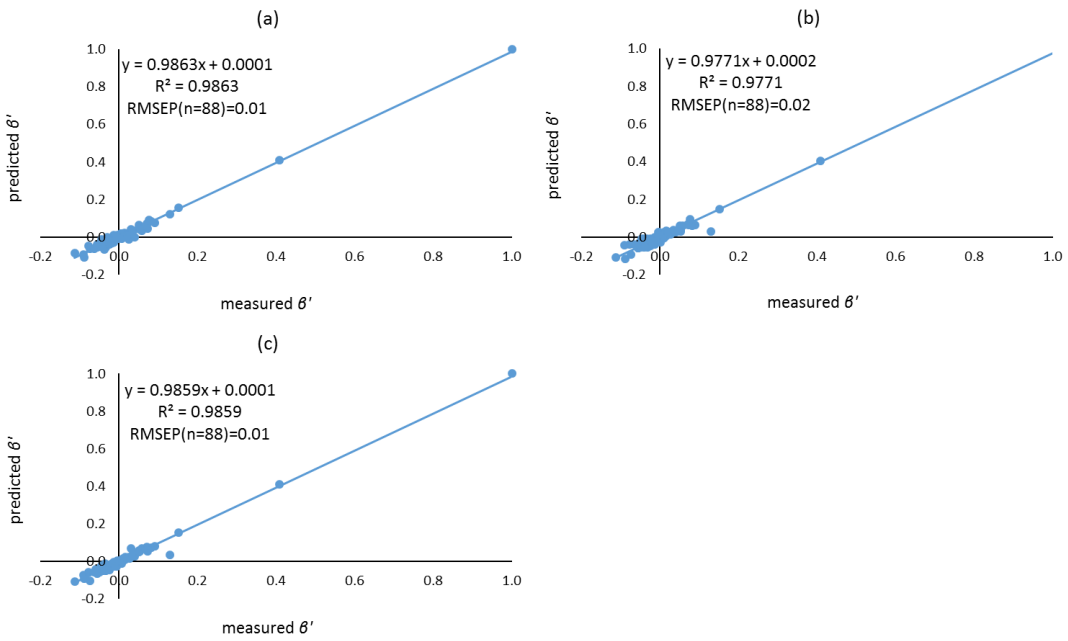


Figure 4.18. Solute coefficient (β') prediction using the LSDI approach with (a) Dragon descriptors, (b) VolSurf+ descriptors, and (c) combined descriptors.

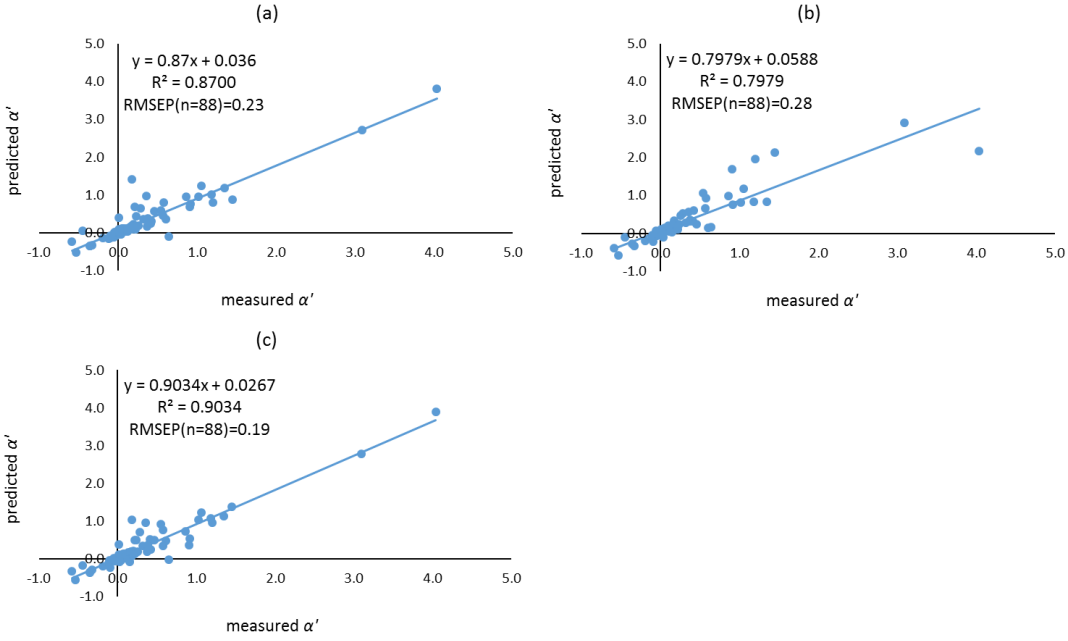


Figure 4.19. Solute coefficient (α') prediction using the LSDI approach with (a) Dragon descriptors, (b) VolSurf+ descriptors, and (c) combined descriptors.

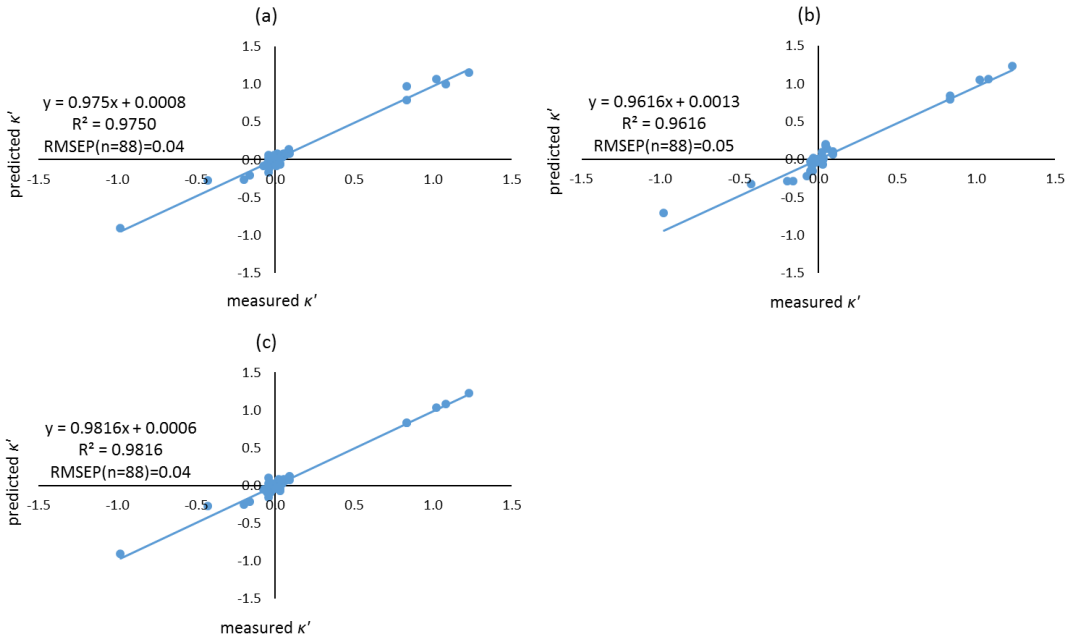


Figure 4.20. Solute coefficient (κ') prediction using the LSDI approach with (a) Dragon descriptors, (b) VolSurf+ descriptors, and (c) combined descriptors.

4.3.2 Performance comparison of filtering approaches

Table 4.3 shows the comparison of the prediction of five solute coefficients using the above-mentioned filtering approaches. As can be seen, in terms of the prediction of solute coefficients, the LSDI approach shows comparable performance to the LCT approach. Results of prediction using the LOO and the Global approach are not acceptable considering the poor correlation coefficients and large errors of prediction. It is worth pointing out that the LCT approach showed the best performance in the prediction of α' and κ' coefficients. The best values of R^2 and RMSEP for α' coefficient prediction obtained from the LCT approach are 0.9536 and 0.14, respectively, which are superior to the values of 0.9034 and 0.19 using the LSDI approach. Similarly, in terms of κ' coefficient prediction, the LCT approach generated R^2 = 0.9907 and RMSEP = 0.03, slightly better than the LSDI approach which gave R^2 = 0.9816 and RMSEP = 0.04.

Table 4.3 shows that the makeup of the training sets is a prime factor in determining the accuracy of the resulting QSRR models. In terms of the LOO and the Global approach, more diverse compounds were employed to build the models, while less diverse compounds were used in the LCT and the LSDI approaches. When compared to the LCT approach, the LSDI approach involved filtering compounds based on the second dominant interaction after hydrophobicity, thus more groups of compounds were generated (six groups). Compounds in each group were representative of a certain type of interaction between compounds and stationary phase, which is more specific than the LCT approach where only three groups were obtained.

Table 4.3. Parameters of prediction of five solute coefficients using different approaches

SCs	MDs	LOO		Global		LCT		LSDI	
		R^2	RMSEP	R^2	RMSEP	R^2	RMSEP	R^2	RMSEP
	Dragon	0.69	0.28	0.65	0.28	0.82	0.21	0.94	0.12
η'	VolSurf+	0.64	0.30	0.83	0.20	0.90	0.16	0.99	0.06
	combined	0.69	0.28	0.67	0.23	0.91	0.15	0.98	0.06
	Dragon	0.40	0.44	0.54	0.34	0.83	0.23	0.92	0.16
σ'	VolSurf+	0.34	0.48	0.61	0.30	0.87	0.20	0.86	0.21
	combined	0.39	0.46	0.71	0.27	0.87	0.21	0.93	0.14
	Dragon	0.06	0.13	0.22	0.05	0.43	0.09	0.99	0.01
β'	VolSurf+	0.08	0.12	0.24	0.04	0.45	0.09	0.98	0.02
	combined	0.07	0.14	0.66	0.03	0.91	0.04	0.99	0.01
	Dragon	0.22	0.56	0.22	0.14	0.93	0.17	0.87	0.23
α'	VolSurf+	0.30	0.54	0.11	0.19	0.95	0.14	0.80	0.28
	combined	0.27	0.55	0.39	0.13	0.95	0.14	0.90	0.19
	Dragon	0.53	0.18	0.78	0.02	0.86	0.10	0.98	0.04
κ'	VolSurf+	0.47	0.21	0.71	0.02	0.99	0.03	0.96	0.05
	combined	0.45	0.20	0.79	0.02	0.99	0.03	0.98	0.04

Furthermore, the performance of retention prediction using the five predicted solute coefficients generated from these filtering approaches was also compared. To achieve that, the predicted five solute coefficients combined with the corresponding five column parameters were fitted into the HSM to calculate compound retention (expressed as the selectivity coefficient (α) relative to ethylbenzene), as detailed in Figures 4.21 – 4.23. As can be seen, no matter which resource of molecular descriptors employed, the LSDI approach showed the best correlation between measured α and predicted α with prediction errors of 0.12, 0.09 and 0.07 for Dragon descriptors, Volsurf+ descriptors, and combined descriptors, respectively.

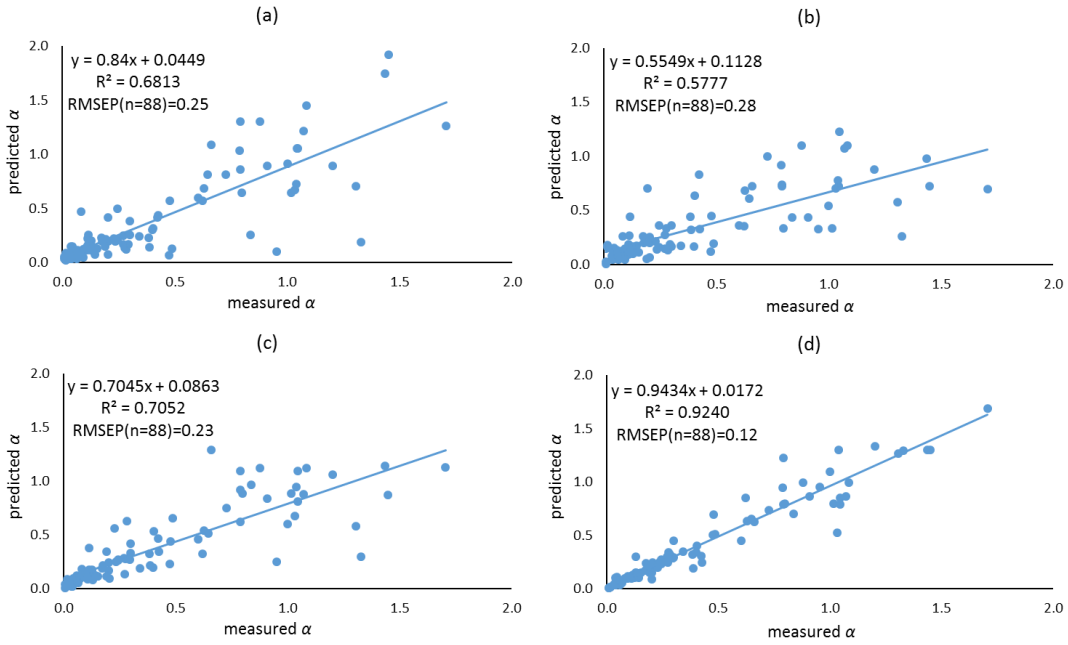


Figure 4.21. Predicted selectivity coefficients ($\alpha=k/k_{EB}$) *via* the full HSM using (a) the LOO, (b) the Global, (c) the LCT, and (d) the LSDI approach with Dragon descriptors.

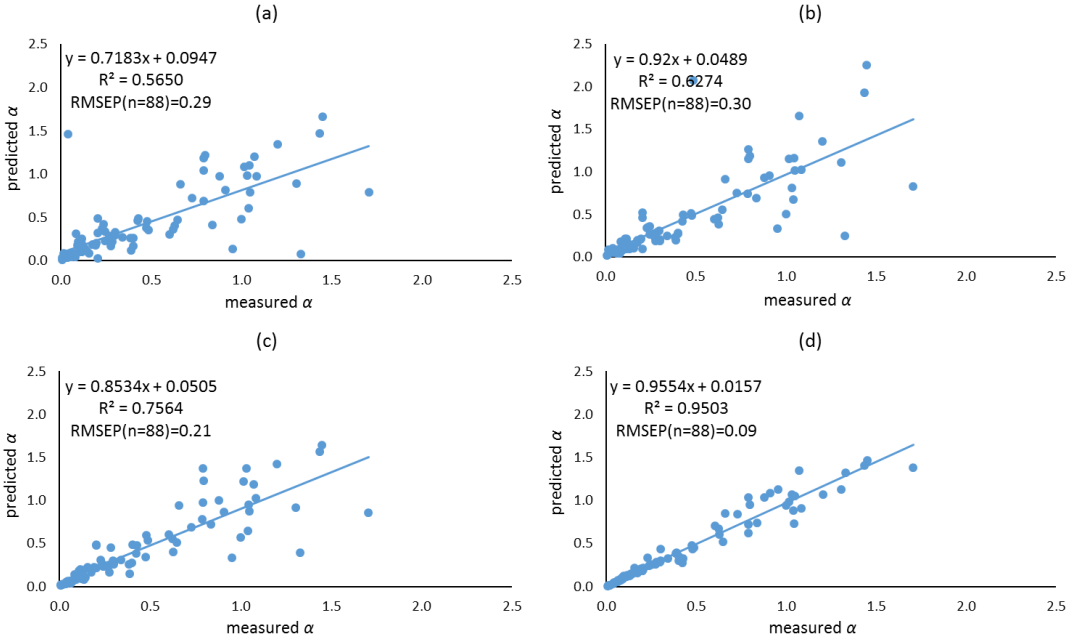


Figure 4.22. Predicted selectivity coefficients ($\alpha=k/k_{EB}$) *via* the full HSM using (a) the LOO, (b) the Global, (c) the LCT, and (d) the LSDI approach with VolSurf+ descriptors.

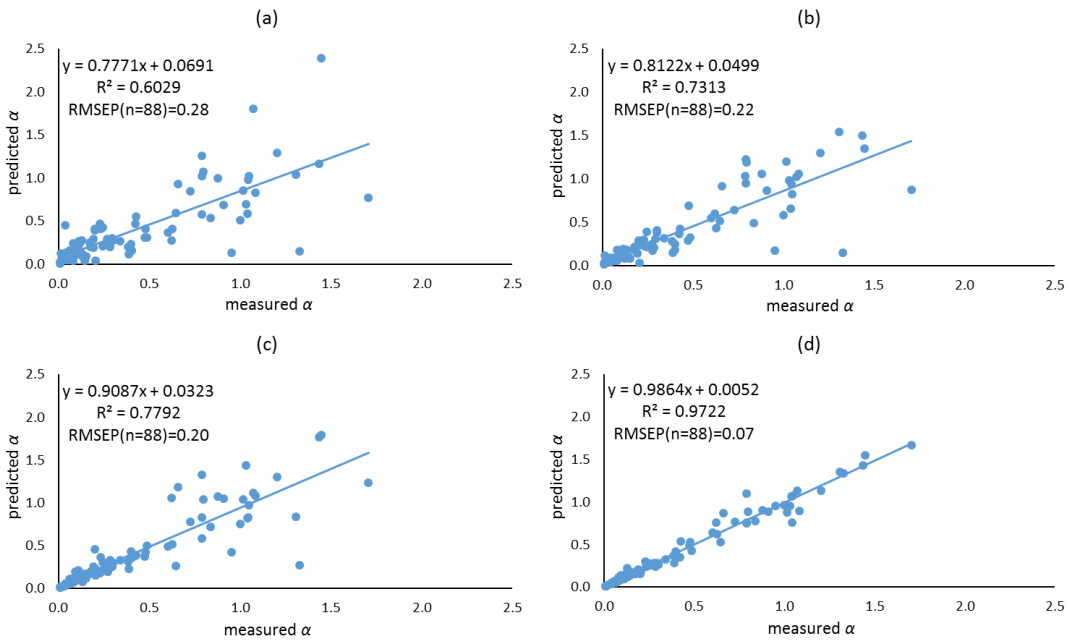


Figure 4.23. Predicted selectivity coefficients ($\alpha=k/k_{EB}$) *via* the full HSM using (a) the LOO, (b) the Global, (c) the LCT, and (d) the LSDI approach with combined descriptors.

4.3.3 Significance of hydrophobicity term in the HSM

There are five terms in the HSM and each term is representative of a certain type of interaction between the compound and the stationary phase. The hydrophobic interaction is well-known for being the primary contributor to the retention in RPLC, as described by the solvophobic theory [19]. The HSM tries to quantify contributions to separation selectivity from other weaker interactions, so it would be interesting to see the relative importance of these interactions regarding their contribution to retention. To this end, the above predicted retention (using predicted solute coefficients) obtained from either Eq. 4.1 using five experimental solute coefficients (*i.e.*, the full HSM), or alternatively Eq. 4.2 using only the hydrophobicity coefficient η' (which we describe here as the approximate HSM) was plotted against the measured retention for 88 compounds in Dataset 1.

$$\log \alpha \equiv \log\left(\frac{k}{k_{\mathrm{EB}}}\right) \approx \eta' \mathbf{H} \qquad 4.2$$

The five predicted solute coefficients modelled using different filtering approaches, combined with the corresponding column coefficients were fitted into Eq. 4.1 (Figures 4.21 – 4.23). Then, the predicted η' values were extracted and fitted into Eq. 4.2 with the same column coefficients for the calculation of retention, Figures 4.24 – 4.26. The approximate HSM using only the hydrophobicity term generated comparable predictions to those obtained using the full HSM where five terms were involved. This indicated that the hydrophobicity term (η' **H)** gives the most important contribution to the retention among the five terms.

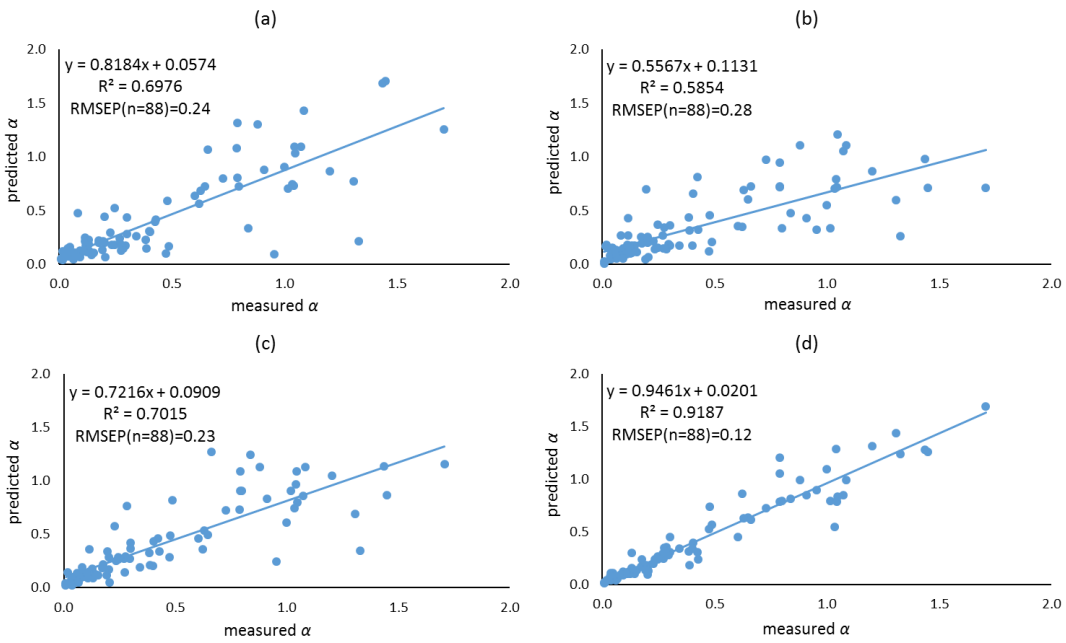


Figure 4.24. Predicted selectivity coefficients ($\alpha=k/k_{EB}$) *via* the approximate HSM using (a) the LOO, (b) the Global, (c) the LCT, and (d) the LSDI approach with Dragon descriptors.

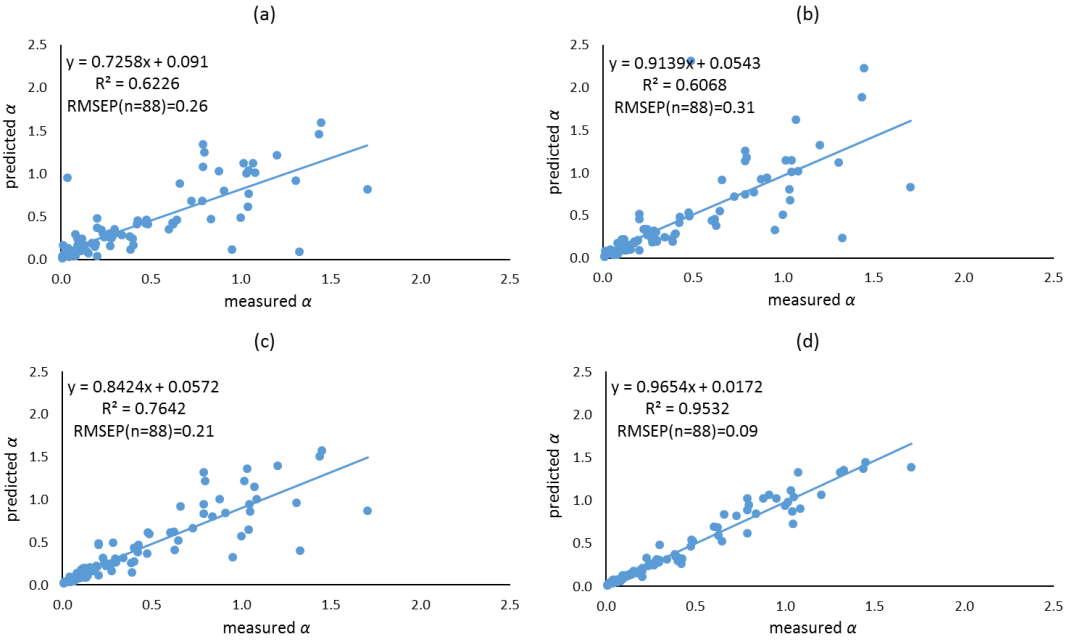


Figure 4.25. Predicted selectivity coefficients ($\alpha=k/k_{EB}$) *via* the approximate HSM using (a) the LOO, (b) the Global, (c) the LCT, and (d) the LSDI approach with VolSurf+ descriptors.

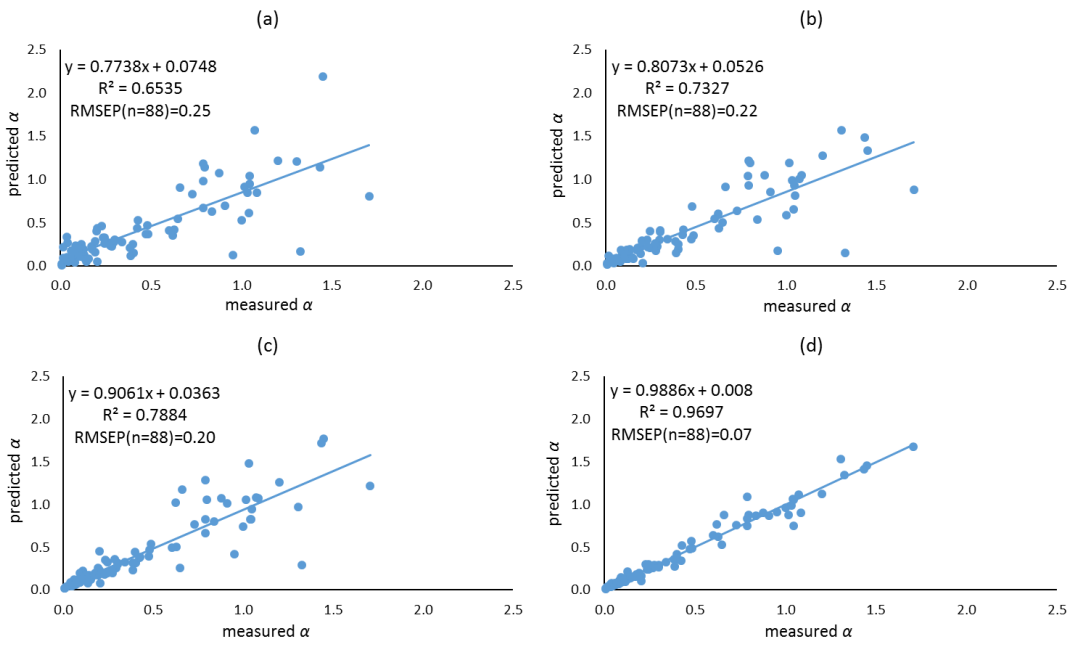


Figure 4.26 Predicted selectivity coefficients ($\alpha=k/k_{EB}$) *via* the approximate HSM using (a) the LOO, (b) the Global, (c) the LCT, and (d) the LSDI approach with the combined descriptors.

Additionally, the calculated retention time (using measured solute coefficients) using both Eq. 4.1 and 4.2 was also plotted against the measured retention time for 88 compounds. Very high correlations were obtained in both cases (Figure 4.27), however the approximate HSM resulted in three times higher error compared with the full HSM (RMSEP of 5.4s *vs.* 1.8s, respectively). Nevertheless, the accuracy of the approximate HSM is sufficient for prediction of the likelihood of co-elution of the compounds. Therefore, for the combined dataset, we have focused on modelling only the hydrophobicity coefficient (η') for each target compound and the subsequent use of the approximate HSM (Eq. 4.2) for the prediction of retention times. Given the fact that almost identical results were obtained when different resources of molecular descriptors used, and that VolSurf+ generated many fewer descriptors, therefore taking less time and effort for calculations, VolSurf+ descriptors were chosen for the modelling of the combined dataset (148 compounds).

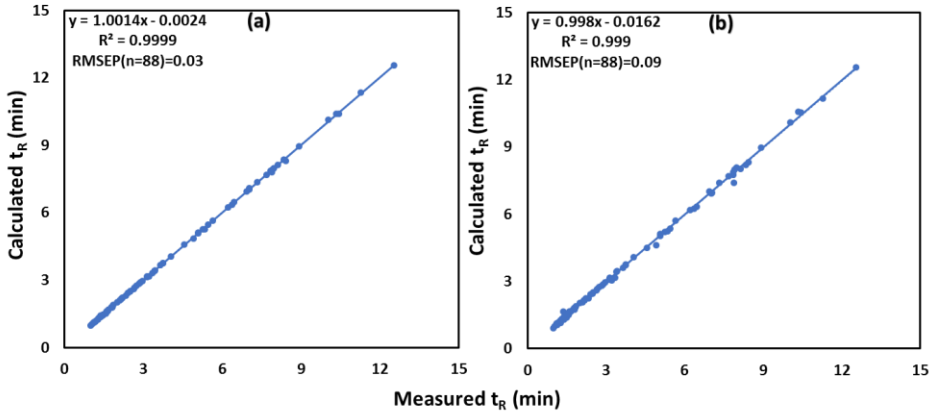


Figure 4.27. Measured retention time *versus* calculated retention time using (a) five experimental solute coefficients *via* the full HSM and (b) experimental hydrophobicity coefficient (η') only *via* the approximate HSM.

4.3.4 QSRR prediction for the combined dataset using the approximate HSM

As a first step, a global PLS model (using a training and test set) was derived using all 126 generated VolSurf+ descriptors for the combined dataset. We will refer to this model as Global 126 (G126) to indicate that all descriptors were used for modelling and that one model was generated to be applied to all compounds in the test set. Figure 4.28a illustrates the correlations obtained between experimental and predicted η' for the 148 compounds. However, pairing the predicted η' values with the column coefficient **H** resulted in an %RMSEP of 24.27 in retention time for the external test set, with errors being generally more pronounced for the longer retained compounds (Figure 4.29a). Additionally, almost identical plots were obtained (Figure 4.30) upon using the full HSM (*i.e.*, by fitting the predicted values of η' obtained from the G126 model but experimental values of other four coefficients into Eq. 4.1) and the approximate HSM (*i.e.*, by fitting the predicted values of η' obtained from the G126 model into Eq. 4.2) for compounds in Dataset 1. This confirmed that the errors obtained for the reference global modelling approach cannot be attributed to the use of the approximate HSM.

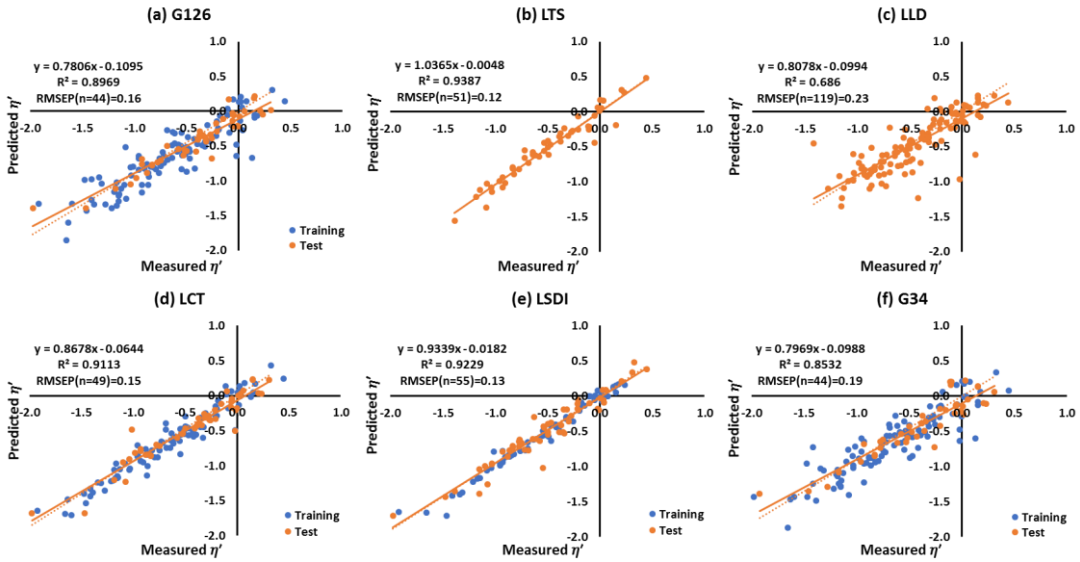


Figure 4.28. Measured η' *versus* predicted η' using (a) G126 (n=44), (b) LTS (TS ≥ 0.5, n=51), (c) LLD (log D ratio < 1.1, n=119), (d) LCT (n=49), (e) LSDI (TS ≥ 0.5, n=55) and (f) G34 (n=44). RMSEP of test set is shown, n is the number of test compounds modelled in each approach. Training and test compounds are highlighted as blue and orange, respectively.

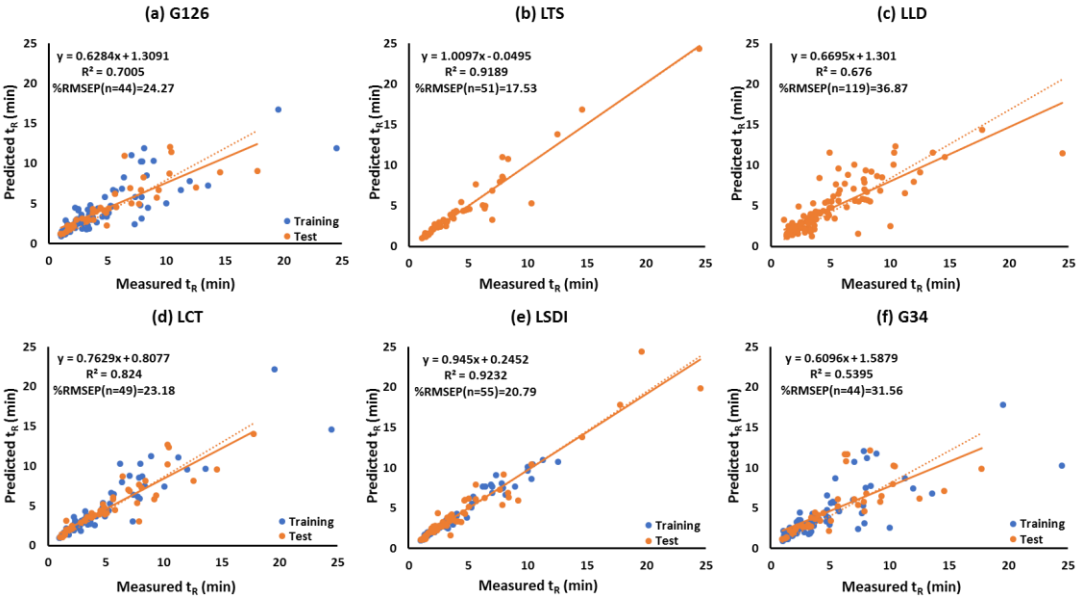


Figure 4.29. Measured retention time *versus* predicted retention time *via* the approximate HSM using (a) G126 (n=44), (b) LTS (TS ≥ 0.5, n=51), (c) LLD (log D ratio < 1.1, n=119), (d) LCT (n=49), (e) LSDI (TS ≥ 0.5, n=55) and (f) G34 (n=44). %RMSEP of test set is shown, n is the number of test compounds modelled in each approach. Training and test compounds are highlighted as blue and orange, respectively.

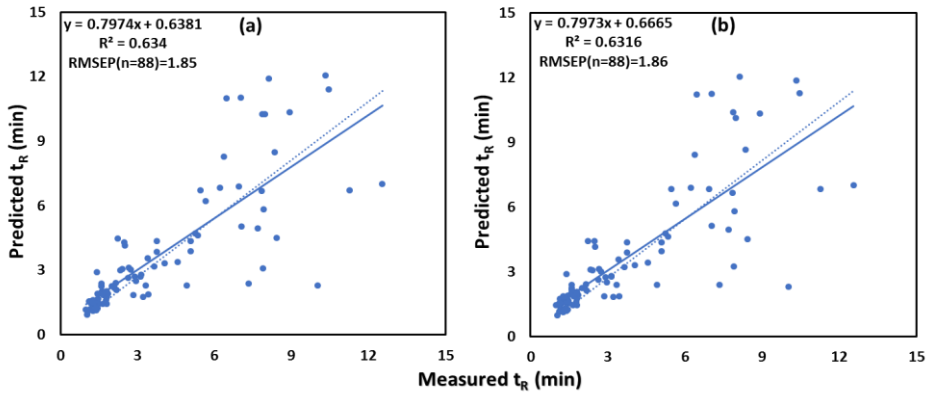


Figure 4.30. Comparison of retention time prediction using (a) predicted η' only obtained from the G126 model *via* the approximate HSM and (b) predicted η' obtained from the G126 model and other four measured solute coefficients *via* the full HSM.

Some conclusions can be obtained based on the relatively high errors obtained for the G126 model and the previous results of the prediction for Dataset 1. First, it is known that QSRR models derived from using a large pool of descriptors often benefit from employing a suitable variable selection technique to identify only the most relevant descriptors to improve the prediction accuracy [9, 13]. Second, previous results have shown that improvement in prediction accuracy can be achieved following compound classification within the dataset to identify the most appropriate set of compounds to yield the QSRR model. Third, local models often provide better performance of prediction than global models [35, 36].

A simple approach for filtering compounds in a dataset can be based on structural similarities derived from the 2D chemical structure of compounds, using the Tanimoto similarity index (the LTS approach). Alternatively, filtering based on a physico-chemical property, such as log D (the LLD approach) can be used. Log D was preferred to log P in this study in order to take into account the ionisation state of the target compound [8, 37]. In implementing either of these two approaches at least five compounds fulfilling the assigned criterion for similarity were required for the training set to make a given target compound eligible for modelling. As can be seen from Figure 4.28b and 4.28c, predictions of η' obtained using the LTS approach were superior to those for the LLD approach, with stronger correlation (R^2 = 0.9387 *vs.* 0.6860). However, only 51 (out of 148) compounds could be modelled using the LTS approach, compared with 119 using the LLD approach, because in many cases a training set of 5 compounds having TS $\geq$ 0.5 could not be identified. Additionally, in a manner similar to the Global model (G126), the accuracy of retention time prediction in both approaches appeared to be markedly higher for compounds that were eluted early (Figure 4.29b and 4.29c). Nevertheless, the average prediction error obtained using the LTS approach translated to about 70 s in retention time, which is reasonably promising to facilitate early prediction of co-elution of compounds, provided there are sufficient similar compounds (with

pairwise TS index of at least 0.5) to each new target in the dataset. Previous results have also showed that even smaller prediction errors can be expected for similarity thresholds greater than 0.5 [8, 13].

As an alternative to the localisation of modelling by deriving one model for each individual eligible compound in the dataset, one model can be derived for each group of similar compounds in the database. The immediate benefit of this approach is a reduced demand for computational resources since fewer QSRR models need to be derived. It has been shown in section 4.2.3 that a simple classification approach could be based on the type (or chemical nature) of compounds (*i.e.* division of compounds in the database into acids, bases or neutrals) [3]. This approach is referred to as the LCT approach. As can be seen in Figure 4.28d and 4.29d, the overall accuracy of using three QSRR models derived from these three clusters of compounds appears to be notably better than the Global model (G126) derived without filtering, being evident from smaller prediction errors and relatively lower scatter of data points. The prediction accuracy given by this method (%RMSEP = 23.18) is also better than the LLD approach (%RMSEP = 36.87) but is just slightly inferior to the LTS approach (%RMSEP = 17.53). However, the advantage of the LCT approach is the simplicity of compound classification as compound groups are made simply by inspecting the chemical structure of compounds and calculating pK_a values.

LSDI	ID	LCT	LSDI	ID	LCT	LSDI	ID	LCT	LSDI	ID	LCT
η' cluster	1	neutral	σ' cluster	31	neutral	α' cluster	55	acid	β' cluster	44	neutral
	2	neutral		32	neutral		56	acid		79	neutral
	3	neutral		33	neutral		57	acid		87	base
	4	neutral		34	neutral		59	acid		88	base
	5	neutral		35	neutral		60	acid	cluster 6	9	neutral
	6	neutral		36	neutral		61	acid		13	neutral
	7	neutral		37	neutral		62	acid		14	neutral
	8	neutral		38	neutral		63	acid		15	neutral
	10	neutral		39	neutral		64	acid		19	neutral
	11	neutral		42	neutral		80	neutral		22	neutral
	12	neutral		43	neutral		81	neutral		25	neutral
	16	neutral		69	neutral		82	neutral		29	neutral
	17	neutral		70	neutral		83	neutral		41	neutral
	18	neutral		71	neutral		84	acid		50	base
	20	neutral		72	neutral		85	neutral		51	base
	21	neutral		73	neutral		86	neutral		52	base
	23	neutral		74	neutral	κ' cluster	45	base		53	base
	24	neutral		75	neutral		46	base		54	base
	26	neutral		76	neutral		47	base		58	acid
	27	neutral		77	neutral		48	base		65	acid
	28	neutral		78	neutral		49	base		66	acid
	30	neutral									
	40	neutral									
	67	neutral									
	68	neutral									

Figure 4.31. Comparison between the LCT (right of compound ID) and the LSDI (left of compound ID) approaches in allocating compounds.

Results obtained for the LCT approach suggested that more localisation of modelling, by identifying more clusters from the dataset, might improve the overall prediction accuracy. As described previously, a compound classification based on the secondary dominant interaction can be used as an underlying premise for a filtering approach (the LSDI approach). Figure 4.31 summarises a comparison between the LCT and the LSDI approaches in allocating compounds to different clusters. As can be seen, most neutral compounds are distributed almost equally between η' and σ' clusters. Two other neutral compounds together with two weak bases create the β' cluster, while the α' cluster comprises both neutral compounds and undissociated acids. Strong bases create the κ' cluster, whereas the remaining weak bases together with some acids and neutrals were not assigned to any cluster in the original study, based on the conclusions that their retention was affected by more than one secondary interaction and they could not be allocated with confidence to a specific cluster [15]. These compounds comprise cluster 6 of the LSDI approach. As can be seen from Figure 4.28e and 4.29e, considerable improvement in prediction accuracy compared to the Global method was again achieved using this approach. More specifically, the much lower deviation of data points from the correlation line and an %RMSEP of 21.91 in retention time prediction was obtained for 27 test compounds from Dataset 1 (Figure 4.32a). Additionally, 28 new test compounds (out of 57) from Dataset 2 that were able to be modelled following the restrictions imposed using this approach (*i.e.*, TS $\geq$ 0.5), exhibited excellent prediction accuracy for all but the two most retained compounds (%RMSEP = 19.66, Figure 4.32b). It is to be hoped that with a larger database containing more compounds with longer retention times, the overall error would be diminished due to increased similarity between compounds. Also, as can be seen in Figure 4.33, while there does not appear to be any strong correlation between absolute prediction error in η' and pairwise TS score, the magnitude of the error remains consistent only at similarities greater than 0.5. This supports previous studies suggesting a TS threshold of 0.5 as a minimum for a reliable prediction in TS-based QSRR modelling [8, 13].

Residual plots of QSRR models using different approaches have been created. Figure 4.34 shows the residuals on the vertical axis and the independent variable (measured t_R) on the horizontal axis. The residual plots for the Global models, and the LLD model, are reasonably evenly distributed around the x-axis with a slight bias towards higher retention times. The LCT shows that early eluted compounds are somewhat better predicted than later eluted compounds. However, the LTS and the LSDI have a more even distribution of residuals, showing a better overall prediction.

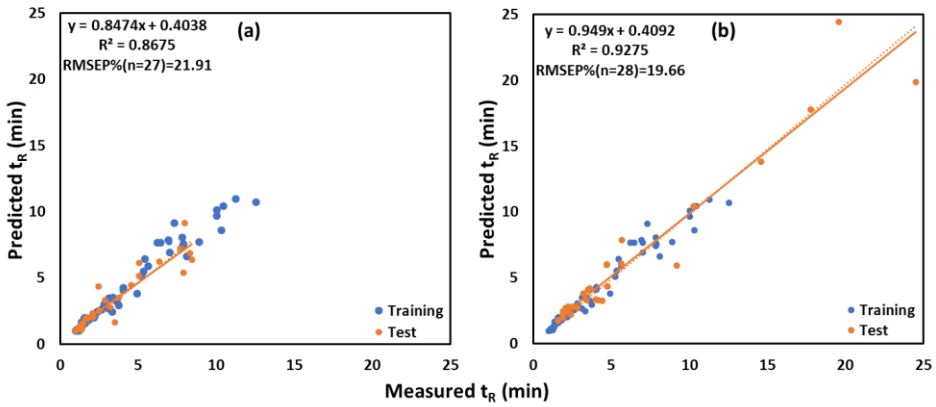


Figure 4.32. Retention prediction using the LSDI approach *via* the approximate HSM for (a) 27 test compounds from Dataset 1 and (b) 28 new test compounds from Dataset 2. %RMSEP of test set is shown, n is the number of test compounds modelled. Training and test compounds are highlighted as blue and orange, respectively.

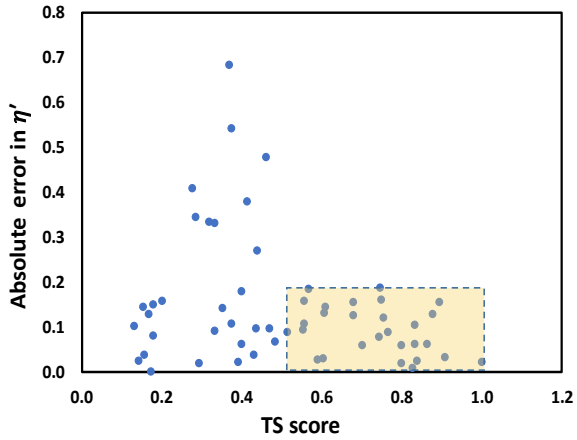


Figure 4.33. Correlation between absolute error in η' and pairwise Tanimoto similarity (TS) score.

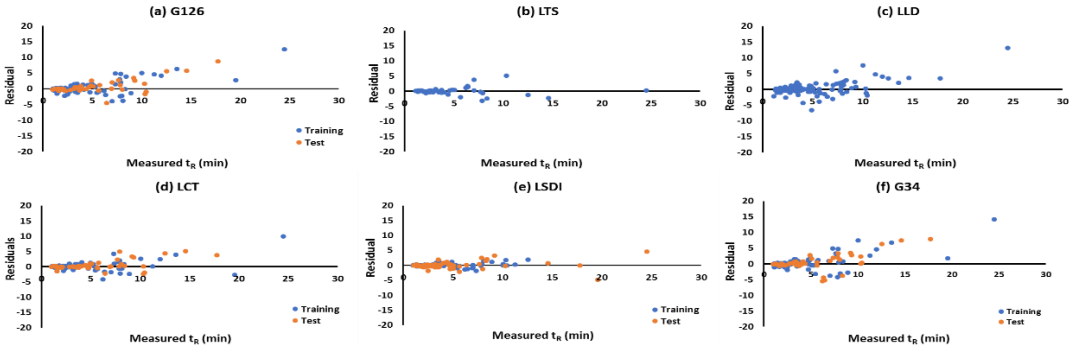


Figure 4.34. Residual plots of QSRR models for retention prediction using different filtering approaches. Training and test compounds are highlighted as blue and orange, respectively.

Table 4.4. Summary of descriptors selected from the LSDI approach

Clusters	Descriptors
η'	MetStab' 'WO1' 'WO2' 'WO3' 'LgD9' 'DRDRDO' 'CW1' 'LgD8' 'CD2'
σ'	WO1' 'WO2' 'LOGP c-Hex' 'L3LgS' 'DRACDO' 'SKIN' '%FU10' 'MetStab' 'LgD7'
β'	DD4' 'WN1' 'W3' 'D1' 'LOGP c-Hex' '%FU5' 'D2' 'WN2' 'SKIN'
α'	LgD10' 'LgD9' 'PB' 'LOGP c-Hex' 'LgD5' 'DD1' 'HSA' 'AUS7.4'
κ'	S' 'G' '%FU7'
cluster 6	LgD5' '%FU8' '%FU7' 'L4LgS' 'LOGP c-Hex' 'L1LgS'

Further examination of the six QSRR models derived following the LSDI approach revealed that 34 descriptors were selected by at least one of these models (Table 4.4). This suggested that the reduction in error could be due to the selection of specific descriptors rather than the local modelling. Therefore, another global PLS model was derived using only these 34 selected descriptors and 70% of the combined dataset as the training set, and was tested using the remaining 30% of compounds as the test set. To avoid ambiguous terminology, this model is called Global 34 (G34). As can be seen in Figure 4.28f and 4.29f, the prediction accuracy of the G34 was far inferior to the overall accuracy of the six local models (Figure 4.28e and 4.29e) despite using the same set of significant descriptors. This comparison reveals the significant advantage of localisation of modelling following an appropriate filtering approach over the global counterpart as the classical approach in QSRR modelling.

4.3.5 Regression Error Characteristics

REC curves provide a valid comparison of the performance of regression models by plotting the error tolerance against the percentage of data points predicted within the tolerance, which is here called the prediction frequency. One can quickly assess the relative merits between various regression models and with the null model by examining the relative position of their REC curves [30].

Figure 4.35 illustrates the REC curves obtained for all the derived models, as well as the null (baseline) model and the ideal model (obtained using the approximate HSM and experimental (rather than predicted) η' values). As can be seen, all the developed QSRR models are significantly more predictive than the null model but are inferior to the ideal model, which exhibits prediction errors less than 30s for nearly 95% of the compounds. In comparison, QSRR models obtained following LSDI, LCT and LTS approaches give similar performance to each other, with a 30s prediction error being achieved for about 70% of the test compounds for each approach. Using a 60s prediction error as criterion, the LSDI approach was slightly superior to the LCT and LTS approaches, with less than 60s prediction error being observed for about 83% of test compounds. As expected, the remaining models, *i.e.* the models derived

from the LLD approach, and the Global models derived using 126 and 34 selected descriptors, were all clearly inferior to the local models and exhibited a prediction accuracy of less than 30s for only 50% of the test compounds which could be modelled.

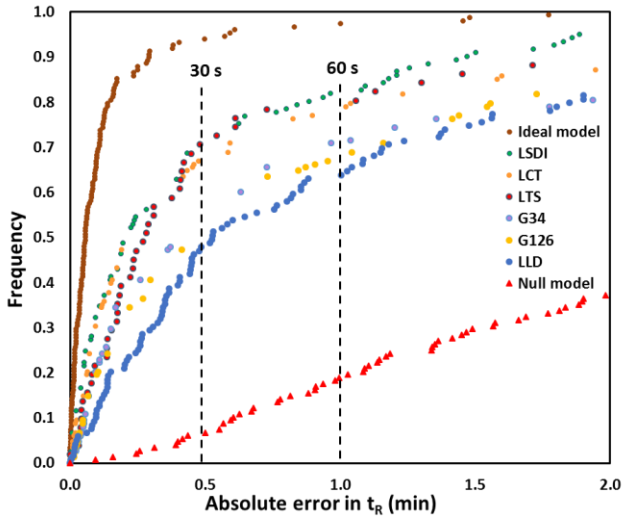


Figure 4.35. REC curves of test compounds comparing the performance of derived models with the null model and the ideal model. The null model was obtained using the mean of the dependent variable (response) as a naïve predicted value for all compounds. The ideal model was obtained using the approximate HSM and experimental (rather than predicted) η' values.

4.3.6 Sum of Ranking Difference analysis

Another overall performance comparison of the proposed approaches can be undertaken using sum of ranking difference (SRD) analysis which has been used for comparison of QSRR models, as well as for other diverse fields [32]. In this work, objects including R^2, Q^2, the %RMSEP of the test set, and the slope of the regression line were selected and evaluated (Table 4.5).

Table 4.5. Input matrix with nine objects and six variables (models) for the Sum of Ranking Difference (SRD) analysis

Analyte	G126	G34	LTS	LLD	LCT	LSDI	Gold Standard
R^2	0.7005	0.5395	0.9189	0.676	0.824	0.9232	0.9943
Q^2	0.6844	0.5523	0.9100	0.6697	0.8173	0.9234	0.9922
slope	0.6284	0.6096	1.0097	0.6695	0.7629	0.9450	1.0000
RMSEP%	24.27	31.56	17.53	36.87	23.18	20.79	4.37
no. of test compounds	44	44	51	119	49	55	148
no. of descriptors used	126	34	45	22	7	7	7
no. of models constructed	1	1	51	119	3	6	1
Prediction frequency with absolute error < 30s	0.49	0.49	0.49	0.67	0.71	0.71	0.94
Prediction frequency with absolute error < 60s	0.68	0.71	0.79	0.64	0.79	0.83	0.97

Moreover, as the applicability of each proposed approach is also important, the number of test compounds, the average number of descriptors used for modelling, the number of constructed models for each approach and the prediction frequency with absolute error less than 30s and 60s based on the REC curves were also taken into account. A given benchmark generated from the ideal correlation obtained using the approximate HSM and experimental η' values was used as the gold standard.

An overview of the ranking and probability level can be seen from Figure 4.36. The LSDI and LCT approach were ranked closest to the gold standard, which demonstrated that these two approaches provided slightly better performance than others and were practically indistinguishable from each other. The LTS approach was ranked third, but still superior to the Global models and the LLD approach. It is not surprising to see that the LLD approach ranked last given the prediction accuracy that was obtained. However, the SRD values for all models were still smaller than those for 95% of models created from randomly ranked numbers, indicating that all proposed approaches can be considered to be predictive. While the LSDI approach provided high prediction accuracy, the allocation of new target compounds to LSDI clusters requires experimental work, which dramatically limits the advantages of the approach. The LCT approach is therefore a more practical and straightforward method, since it is easy to allocate new compounds to clusters. Furthermore, for the databases used in this study, the LCT approach could be applied to 148 compounds, compared to 121 for the LSDI and 51 for the LTS approaches.

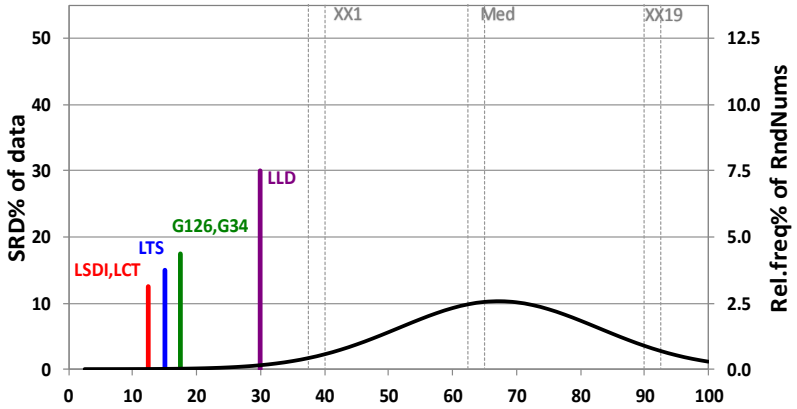


Figure 4.36. SRD-CRRN test results of the data matrix given in Table 4.5. SRD values are plotted on *x* axis and left *y* axis, right *y* axis shows the relative frequencies (%).

4.3.7 Molecular descriptors

One distinctive feature of VolSurf+ descriptors is their versatility and ease of interpretation given that their calculation is based on generic physico-chemical properties of molecules [38]. Another advantage of VolSurf+ descriptors over other descriptor generator platforms is that all the prerequisite steps before descriptor calculations, including 3D

structure generation, conformational analysis and geometry optimisation are performed automatically by the software, making descriptor calculation both user- and platform-independent and therefore less prone to error. In addition, the number of generated descriptors is much less than other platforms, so a descriptor selection method like GA-PLS can be performed simply and rapidly.

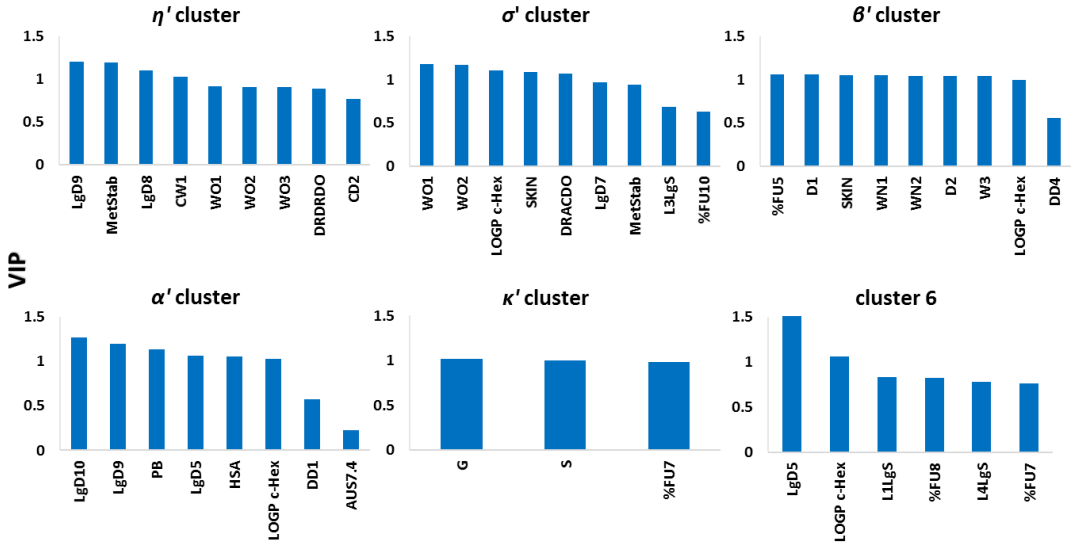


Figure 4.37. PLS variable importance to projection for the optimised GA-PLS models generated for the LSDI model.

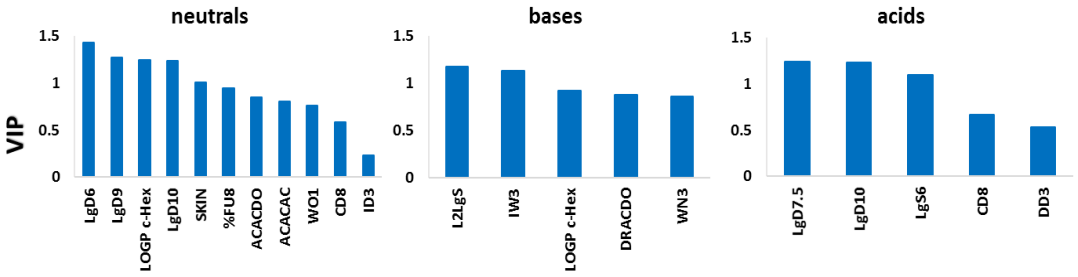


Figure 4.38. PLS variable importance to projection for the optimised GA-PLS models generated for the LCT model.

Table 4.4 summarises the 34 unique descriptors selected by at least one of the six QSRR models derived from the LSDI approach, many of which were also found in common with the LCT approach. The variable importance to projection (VIP) plot (Figures 4.37 – 4.38) shows the relative importance of each descriptor in the optimised GA-PLS models. Not surprisingly, the strongest correlations are observed for log D descriptors. Other significant descriptors with more apparent contribution to the RPLC retention mechanism include log P, H-bond donor volumes (WO), the percentage of unionised species at different pH (%FU), descriptors representing hydrophilicity/hydrophobicity or in direct relationship with them, including the accessible surface area of the molecule by water as a hydrophilic probe (S), molecular globularity (G), which represents both molecular bulkiness and molecular flexibility,

descriptors relating to the hydrophilic volume (W and CW), the hydrophobic volume (D, DD and CD).

4.3.8 Co-elution prediction using the proposed QSRR method

The primary objective in this study of QSRR modelling of retention time was its subsequent use for predicting retention times of compounds based on their chemical structure. Since Dataset 1 involves 88 compounds and 10 reversed-phase columns, it is possible to evaluate the application of retention prediction on a range of columns. Five representative compounds, 2-nitrobenzoic acid, 2,5-dinitrophenol, N-butylaniline, toluene and p-chlorotoluene were selected. It is noteworthy that these test compounds include acidic, basic and neutral compounds. The retention of these compounds was modelled on nine of the ten columns in the dataset as the YMC Pack Pro C18 column (column number 9 in Wilson's work) was excluded since the retention times of 2,5-dinitrophenol and N-butylaniline were not available for that column [14, 15]. Predicted η' coefficients of the five compounds obtained using the LSDI QSRR model were used to predict retention times and Table 4.6 lists the predicted and actual retention times on nine columns. Predicted chromatograms of the five compounds on the nine columns were simulated using Microsoft Excel Macros [39] and two examples of the simulated chromatograms, together with the actual retention time of each compound, are shown in Figure 4.39, with the remaining chromatograms being shown in Figure 4.40. In all cases, the predicted retention times were very close to the actual retention times. The accuracy of retention prediction using QSRR combined with the approximate HSM indicates that this approach is appropriate for determining the potential of co-elution of the target compounds, based only on their chemical structures.

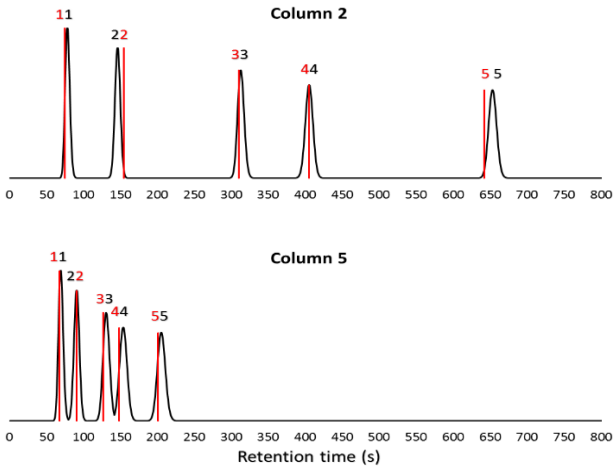


Figure 4.39. Simulated chromatograms of five representative compounds on column 2 (Symmetry) and column 5 (SB-300). The red lines represent the actual retention times. Numbering of representative compounds in graph: 1, 2-nitrobenzoic acid; 2, 2,5-dinitrophenol; 3, *N*-butylaniline; 4, toluene and 5, *p*-chlorotoluene.

Table 4.6. Experimental and predicted retention time (s) of five representative compounds on nine columns (C1 to C8, and C10)

Representative compounds	C 1		C 2		C 3		C 4		C 5	
	Exp	Pred	Exp	Pred	Exp	Pred	Exp	Pred	Exp	Pred
2-nitrobenzoic acid	85	85	75	77	67	69	70	72	68	68
2,5-dinitrophenol	184	171	155	145	129	122	135	127	91	88
N-butylaniline	375	373	310	312	243	245	248	249	127	130
toluene	479	481	405	404	305	311	307	313	148	151
p-chlorotoluene	762	768	642	652	471	484	470	481	201	204
Representative compounds	C 6		C 7		C 8		C 10			
	Exp	Pred	Exp	Pred	Exp	Pred	Exp	Pred		
2-nitrobenzoic acid	66	69	84	86	81	84	81	83		
2,5-dinitrophenol	132	125	168	157	161	152	129	125		
N-butylaniline	265	259	328	322	320	313	218	220		
toluene	329	333	405	411	392	400	269	271		
p-chlorotoluene	515	530	638	645	611	633	401	405		

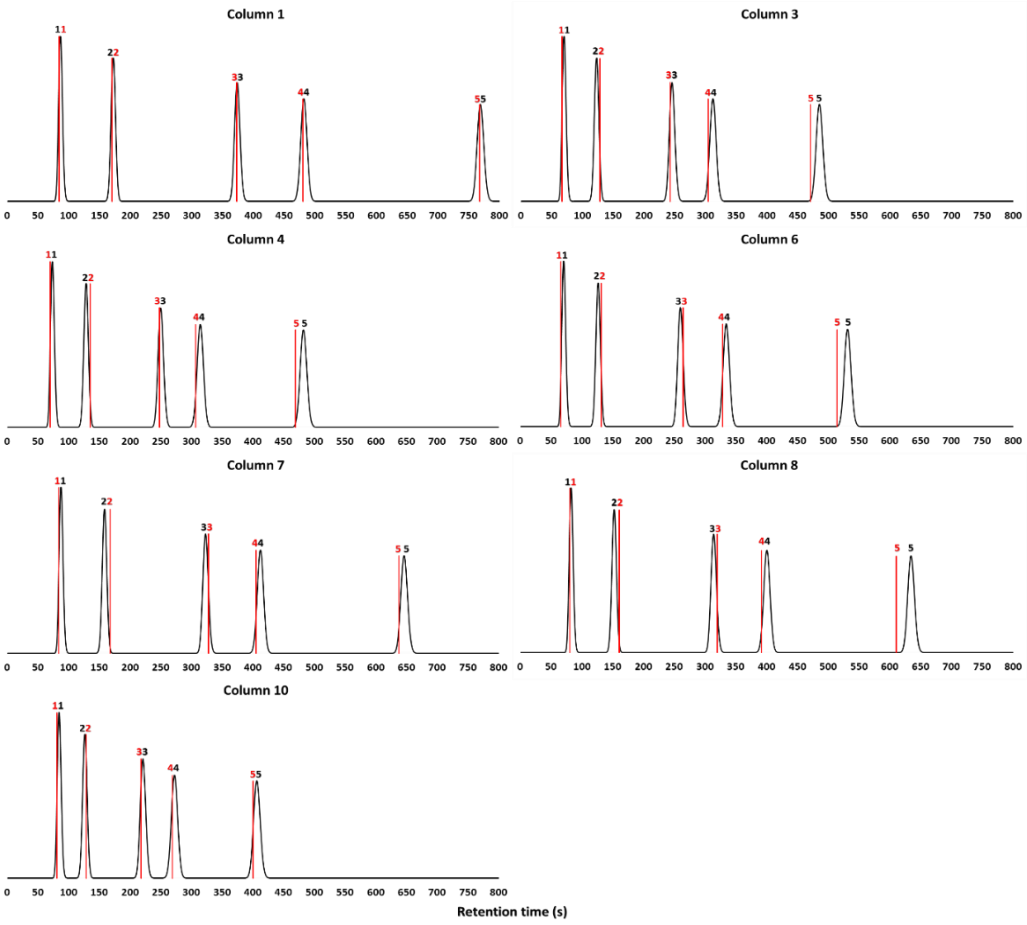


Figure 4.40. Simulated chromatograms of five representative compounds on the remaining seven columns. The red lines represent the actual retention times. Numbering of representative compounds in graph: 1, 2-nitrobenzoic acid; 2, 2,5-dinitrophenol; 3, *N*-butylaniline; 4, toluene and 5, *p*-chlorotoluene. Numbering of columns in graph: 1, Inertsil; 3, SB-100; 4, SB-90; 6, Eclipse; 7, YMC 15; 8, YMC 16 and 10, Discovery.

4.3.9 Retention prediction for new compounds using the proposed QSRR method

Given the robust performance of the constructed QSRR models with the five representative compounds selected from the database, the predictive ability for new compounds that have never been used in the modelling process was also investigated. To achieve this, five analytical grade compounds including an acid, a base and three neutral compounds were selected and purchased from Sigma-Aldrich (St. Louis, MO, USA): pindolol, 1, 8-dihydroxynaphthalene, 4-ethylnitrobeneze, 2-phenylbutane and 4-heptylbenzoic acid. The columns employed in this study were all obtained from Thermo Fisher Scientific, an Acclaim^{TM} 120 C18 column (4.6 × 150 mm, 5.0 μm); an Accucore C18 column (4.6 × 150 mm, 2.6 μm); a Hypersil GOLD C8 column (4.6 × 150 mm, 5.0 μm); a Hypersil GOLD C18 column (4.6 × 150 mm, 5.0 μm) and a Hypersil ODS C18 column (4.6 × 150 mm, 5.0 μm). The HSM coefficients for the columns used in this work were extracted from the USP website. Solute coefficients of η' for the five new compounds were predicted using both the LCT and LSDI approaches. Retention data were collected under the same chromatographic conditions as Wilson's previous work [25]. Uracil was used to determine the void time of the columns. Ethylbenzene was chosen as the reference compound for calculation of predicted retention times.

Before approaching QSRR modelling, the five new compounds were allocated into clusters. For the LCT approach, pindolol, and 4-heptylbenzoic acid were allocated into the base and the acid cluster, respectively, and the other three neutrals were allocated into the neutrals cluster. Unlike the LCT approach, for the LSDI filtering, a Tanimoto similarity analysis was involved to ensure that the new compounds could be allocated into the correct cluster based on structural similarity by finding the nearest similar compound from the previously used database. 1, 8-dihydroxynaphthalene and 2-phenylbutane was allocated into the η' cluster, pindolol was assigned into the κ' cluster, and 4-heptylbenzoic acid and 4-ethylnitrobenzene was grouped into the α' cluster and cluster 6, respectively. Their η' coefficients were therefore predicted using the corresponding clusters in the database as the training sets. Finally, the predicted η' coefficients combined with the available column parameters were fitted into the approximate HSM to yield their retention times, and predicted retention times were compared to the measured retention times.

Table 2.9 in Chapter 2 lists the measured retention times of the five new representative compounds on the five columns used for data collection. Figure 4.41 shows the correlation between the predicted and measured retention times of the five compounds using the LCT and the LSDI approaches, respectively. As can be seen, good correlations were observed for both cases with R^2 of 0.9634 and 0.9861, respectively. In terms of prediction error, the LCT

approach generated a %RMSEP of only 15.41, superior to the LSDI approach with a %RMSEP of 28.67. The powerful performance of the models illustrates the feasibility of using the proposed approaches to predict retention for new compounds. As mentioned previously, compared to the LSDI approach, the LCT is a straightforward method where training sets can be formed easily based on the compound's type. However, the LSDI approach requires further allocation of test compounds to choose the correct training set for modelling, which increases the complexity of the application.

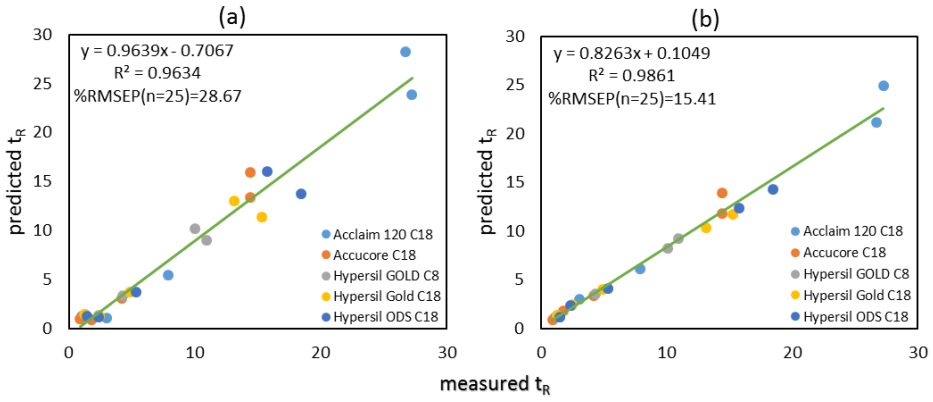


Figure 4.41. Comparison of retention time prediction for five new compounds on five columns using (a) the LSDI approach and (b) the LCT approach *via* the approximate HSM.

4.4 Conclusions

This study has evaluated the feasibility of utilising the HSM and QSRR for predicting retention times in RPLC. This was achieved through modelling only the solute coefficient η' representing compound hydrophobicity in the HSM. It was shown that modelling other solute coefficients in the HSM is not necessary to achieve sufficient prediction accuracy. The resources of molecular descriptors have almost no contribution to the performance of predictive ability for the constructed QSRR models. Different approaches were used for filtering the dataset to derive local models. Results showed improvement in prediction accuracy for the LSDI, LCT and LTS approaches compared with global models derived from the whole dataset without filtering. Of these approaches, the LCT exhibits advantages in its ease of application and the larger number of compounds which can be modelled using this approach. However, the LTS approach was the simplest to apply and provided that there was a sufficient number of similar compounds in the dataset which met the TS score cut-off, it yielded sufficiently accurate results.

It is recognised that modelling based only on η' can predict only the hydrophobic selectivity of compounds and that the predicted elution order of target compounds is determined solely by their η' values. Nevertheless, the predictions reported here show

sufficient accuracy to meet the major objective of this study, namely to determine the likelihood of co-elution of compounds. Future work includes improving the prediction accuracy of the models by including descriptors from other resources, modelling more of the HSM coefficients to better describe selectivity, and using more advanced filtering approaches, such as dual-filtering [9]. The scope of modelling can also be expanded by investigating more columns and including more compounds of greater Tanimoto similarity in the database. With column coefficients being already available for nearly 700 commercial C8 and C18 phases, the proposed approach is anticipated to assist drug development in the pharmaceutical industry and to find application in other industries.

4.5 References

1. Bączek, T., R. Kaliszan, K. Novotná, and P. Jandera, *Comparative characteristics of HPLC columns based on quantitative structure–retention relationships (QSRR) and hydrophobic-subtraction model.* Journal of Chromatography A, 2005. **1075**(1-2): p. 109-115.
2. Kazakevich, Y.V. and R. Lobrutto, *HPLC for pharmaceutical scientists*. 2007: John Wiley & Sons.
3. Muteki, K., J.E. Morgado, G.L. Reid, J. Wang, G. Xue, F.W. Riley, J.W. Harwood, D.T. Fortin, and I.J. Miller, *Quantitative structure retention relationship models in an analytical Quality by Design framework: simultaneously accounting for compound properties, mobile-phase conditions, and stationary-phase properties.* Industrial & Engineering Chemistry Research, 2013. **52**(35): p. 12269-12284.
4. Dove, S., *Multivariate analysis of hydrophobic descriptors.* ADMET and DMPK, 2014. **2**(1): p. 3-17.
5. Talebi, M., S.H. Park, M. Taraji, Y. Wen, R.I.J. Amos, P.R. Haddad, R.A. Shellie, R. Szucs, C.A. Pohl, and J.W. Dolan, *Retention Time Prediction Based on Molecular Structure in Pharmaceutical Method Development: A Perspective.* LCGC North America, 2016. **34**(8): p. 550–558.
6. Baumann, K., *Cross-validation as the objective function for variable-selection techniques.* TrAC Trends in Analytical Chemistry, 2003. **22**(6): p. 395-406.
7. Talebi, M., G. Schuster, R.A. Shellie, R. Szucs, and P.R. Haddad, *Performance comparison of partial least squares-related variable selection methods for quantitative structure retention relationships modelling of retention times in reversed-phase liquid chromatography.* Journal of Chromatography A, 2015. **1424**: p. 69-76.
8. Tyteca, E., M. Talebi, R. Amos, S.H. Park, M. Taraji, Y. Wen, R. Szucs, C.A. Pohl, J.W. Dolan, and P.R. Haddad, *Towards a chromatographic similarity index to establish localized quantitative structure-retention models for retention prediction: Use of retention factor ratio.* Journal of Chromatography A, 2017. **1486**: p. 50-58.
9. Taraji, M., P.R. Haddad, R.I.J. Amos, M. Talebi, R. Szucs, J.W. Dolan, and C.A. Pohl, *Rapid Method Development in Hydrophilic Interaction Liquid Chromatography for Pharmaceutical Analysis Using a Combination of Quantitative Structure–Retention Relationships and Design of Experiments.* Analytical Chemistry , 2017. **89**(3): p. 1870-1878.
10. Wang, C., M.J. Skibic, R.E. Higgs, I.A. Watson, H. Bui, J. Wang, and J.M. Cintron, *Evaluating the performances of quantitative structure-retention relationship models with different sets of molecular descriptors and databases for high-performance liquid chromatography predictions.* Journal of Chromatography A, 2009. **1216**(25): p. 5030-5038.

11. N.S. Wilson, M.D. Nelson, J.W. Dolan, L.R. Snyder, and P.W. Carr, *Column selectivity in reversed-phase liquid chromatography II. Effect of a change in conditions.* Journal of Chromatography A, 2002. **961**: p. 195–215.
12. Park, S.H., R.A. Shellie, G.W. Dicinoski, G. Schuster, M. Talebi, P.R. Haddad, R. Szucs, J.W. Dolan, and C.A. Pohl, *Enhanced methodology for porting ion chromatography retention data.* Journal of Chromatography A, 2016. **1436**: p. 59-63.
13. Park, S.H., P.R. Haddad, M. Talebi, E. Tyteca, R.I. Amos, R. Szucs, J.W. Dolan, and C.A. Pohl, *Retention prediction of low molecular weight anions in ion chromatography based on quantitative structure-retention relationships applied to the linear solvent strength model.* Journal of Chromatography A, 2017. **1486**: p. 68-75.
14. N.S. Wilson, J.W. Dolan, L.R. Snyder, P.W. Carr, and L.C. Sander, *Column selectivity in reversed-phase liquid chromatography III. The physico-chemical basis of selectivity.* Journal of Chromatography A, 2002. **961**: p. 217–236.
15. N.S. Wilson, M.D. Nelson, J.W. Dolan, L.R. Snyder, R.G. Wolcott, and P.W. Carr, *Column selectivity in reversed-phase liquid chromatography I. A general quantitative relationship.* Journal of Chromatography A, 2002. **961**: p. 171–193.
16. D.H. Marchand, P.W. Carr, D.V. McCalley, U.D. Neue, J.W. Dolan, and L.R. Snyder, *Contributions to reversed-phase column selectivity. II. Cation exchange.* Journal of Chromatography A, 2001. **1218**: p. 7110–7129.
17. Carr, P.W., J.W. Dolan, U.D. Neue, and L.R. Snyder, *Contributions to reversed-phase column selectivity. I. Steric interaction.* Journal of Chromatography A, 2011. **1218**(13): p. 1724-1742.
18. Carr, P.W., J.W. Dolan, J.G. Dorsey, L.R. Snyder, and J.J. Kirkland, *Contributions to reversed-phase column selectivity: III. Column hydrogen-bond basicity.* Journal of Chromatography A, 2015. **1395**: p. 57-64.
19. Snyder, L.R., *A New Look at the Selectivity of RPC Columns.* Analytical Chemistry , 2007. **79**(9): p. 3254–3262.
20. Crivori, P., G. Cruciani, P.-A. Carrupt, and B. Testa, *Predicting blood− brain barrier permeation from three-dimensional molecular structure.* Journal of Medicinal Chemistry, 2000. **43**(11): p. 2204-2216.
21. *HPLC Columns*. [cited 2017 July]; Available from: http://www.hplccolumns.org/database/compare.php.
22. LC Tan, PW Carr, and M. Abraham, *Study of retention in reversed-phase liquid chromatography using linear solvation energy relationships I. The stationary phase.* Journal of Chromatography A, 1996. **752**: p. 1-18.
23. Leardi, R. and A.L. Gonzalez, *Genetic algorithms applied to feature selection in PLS regression: how and when to use them.* Chemometrics and Intelligent Laboratory Systems, 1998. **41**(2): p. 195-207.
24. Marengo, E. and R. Todeschini, *A new algorithm for optimal, distance-based experimental design.* Chemometrics and Intelligent Laboratory Systems, 1992. **16**: p. 37-44.
25. Wilson, N., M. Nelson, J. Dolan, L. Snyder, R. Wolcott, and P. Carr, *Column selectivity in reversed-phase liquid chromatography: I. A general quantitative relationship.* Journal of Chromatography A, 2002. **961**(2): p. 171-193.
26. Tyteca, E., M. Talebi, R. Amos, S.H. Park, M. Taraji, Y. Wen, R. Szucs, C.A. Pohl, J.W. Dolan, and P.R. Haddad, *Towards a chromatographic similarity index to establish localized quantitative structure-retention models for retention prediction: use of retention factor ratio.* Journal of Chromatography A, 2017. **1486**: p. 50-58.
27. Wen, Y., M. Talebi, R.I. Amos, R. Szucs, J.W. Dolan, C.A. Pohl, and P.R. Haddad, *Retention prediction in reversed phase high performance liquid chromatography using quantitative structure-retention relationships applied to the Hydrophobic Subtraction Model.* Journal of Chromatography A, 2018. **1541**: p. 1-11.

28. Tropsha, A., P. Gramatica, and V.K. Gombar, *The Importance of Being Earnest-Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models.* QSAR & Combinatorial Science, 2003. **22**: p. 69-77.
29. Searson, D.P., in *GPTIPS 2: an open-source software platform for symbolic data mining*. 2015, Springer. p. 551-573.
30. Bi, J. and K.P. Bennett, *Regression Error Characteristic Curves*, in *Twentieth International Conference on Machine Learning (ICML-2003)*. 2003: Washington DC. p. 43-50.
31. Héberger, K., *Sum of ranking differences compares methods or models fairly.* TrAC Trends in Analytical Chemistry, 2010. **29**(1): p. 101-109.
32. Kollár-Hunek, K. and K. Héberger, *Method and model comparison by sum of ranking differences in cases of repeated observations (ties).* Chemometrics and Intelligent Laboratory Systems, 2013. **127**: p. 139-146.
33. Rácz, A., D. Bajusz, and K. Héberger, *Consistency of QSAR models: Correct split of training and test sets, ranking of models and performance parameters.* SAR and QSAR in Environmental Research, 2015. **26**(7-9): p. 683-700.
34. Héberger, K. and K. Kollár-Hunek, *Sum of ranking differences for method discrimination and its validation: comparison of ranks with random numbers.* Journal of Chemometrics, 2011. **25**(4): p. 151-158.
35. Park, S.H., P.R. Haddad, R.I. Amos, M. Talebi, R. Szucs, C.A. Pohl, and J.W. Dolan, *Towards a chromatographic similarity index to establish localised Quantitative Structure-Retention Relationships for retention prediction. III Combination of Tanimoto similarity index, logP, and retention factor ratio to identify optimal analyte training sets for ion chromatography.* Journal of Chromatography A, 2017. **1520**: p. 107-116.
36. Taraji, M., P.R. Haddad, R.I. Amos, M. Talebi, R. Szucs, J.W. Dolan, and C.A. Pohl, *Use of dual-filtering to create training sets leading to improved accuracy in quantitative structure-retention relationships modelling for hydrophilic interaction liquid chromatographic systems.* Journal of Chromatography A, 2017. **1507**: p. 53-62.
37. Taraji, M., P.R. Haddad, R.I. Amos, M. Talebi, R. Szucs, J.W. Dolan, and C.A. Pohl, *Prediction of retention in hydrophilic interaction liquid chromatography using solute molecular descriptors based on chemical structures.* Journal of Chromatography A, 2017. **1486**: p. 59-67.
38. Randazzo, G.M., D. Tonoli, S. Hambye, D. Guillarme, F. Jeanneret, A. Nurisso, L. Goracci, J. Boccard, and S. Rudaz, *Prediction of retention time in reversed-phase liquid chromatography as a tool for steroid identification.* Analytica Chimica Acta, 2016. **916**: p. 8-16.
39. Kadjo, A. and P.K. Dasgupta, *Tutorial: Simulating chromatography with Microsoft excel macros.* Analytica Chimica Acta, 2013. **773**: p. 1-8.

5 Retention Index Prediction to Improve Structure Identification in Non-Targeted Metabolomics

5.1 Introduction

Targeted and non-targeted strategies are involved in metabolomics studies [1-3]. For the targeted strategy, the known metabolites are often measured quantitatively based on the predefined metabolite-specific signals [2, 3]. In contrast, all the unknown metabolites present in a sample of interest need to be measured in non-targeted metabolomics (NTM) which involves high-throughput and comprehensive analysis. Therefore, additional methods for the subsequent interpretation by means of *in silico* or experimental routines must be employed [2, 4]. For example, analytical platforms like nuclear magnetic resonance spectroscopy (NMR) and mass spectrometry (MS) coupled to liquid chromatography as a separation technique, have enabled rapid detection in NTM for a large number of metabolites [5, 6]. For the non-targeted strategy, many steps are involved including sample selection, sample analysis, data treatment, and metabolite identification ,which is the ultimate goal in NTM, allowing analytical data to be converted into meaningful biological knowledge [7, 8]. However, a confident metabolite identification requires significant effort compared to the targeted strategy given the fact that the unknown metabolites cover a diverse chemical space in NTM [5, 9]. Although advanced strategies like MS/NMR can provide much information to speed up the identification of metabolites, authentic pure chemical standards are still needed for unequivocal identification in NTM [4, 10]. Until now, confident metabolite identification is still a bottleneck because it is often time-consuming, costly, and frequently unsuccessful. The process of metabolite identification in NTM is complex and highly dependent on the robustness of the analytical techniques applied, as well as the databases and resources utilised for mass-based searching [3, 4, 11].

An advantage of the LC-MS-based technique in NTM is that it increases the possibility of metabolite identification with a combination of sensitivity and selectivity [6, 12, 13]. There are several steps of metabolite identification in LC-MS-based NTM analysis. Firstly, an accurate mass of the eluted metabolite is identified based on a target peak displaying metabolomics information. Then, the potential molecular formulae corresponding to the target peak are defined through mass-based searching using electronic resources like the Human Metabolome Database (HMDB), METLIN, LMSD, MassBank, RIKEN and PubChem [1, 8]. Finally, the returned matches based on the exact mass of the metabolite then can be identified and confirmed using additional experimental data. This routine method has been used years for the identification of metabolites in NTM, however, a limitation still exists. For example, a candidate compound corresponding to the exact mass may not be found, or in many cases, an

excessive number of candidates may be returned, which means that eliminating these false positives is necessary [4, 6, 14]. The candidate metabolites often behave similarly with respect to their exact mass but differently in biological systems, which increases the probability of misidentification [5, 15].

Chromatographic retention, a specific property that is readily available for LC-MS-based data, has been used as a feature for the identification of metabolites. For example, Kaliszan's group has used retention comparison to remove false positive identifications during the interpretation of metabolomics data in the area of doping control [4]. Moreover, Aicheler and co-workers have improved the identification rate in non-targeted lipidomic approaches by using a retention time filter to reduce the rate of false positives [1]. There is a strong demand for a tool to predict chromatographic behaviour, and as a result, to identify new analytes and their metabolites [4, 8, 16, 17]. These objectives can be achieved using a combination of analytical routines and chemometric techniques, and quantitative structure–retention relationship (QSRR) modelling is a promising solution from a variety of chemometric methods [4, 18-22].

One of the aims of QSRR is to quantitatively determine the relationships between chromatographic retention parameters and compound structures [20, 23, 24]. Usually, QSRR models are built as functions of molecular descriptors that are generated based on the structures of compounds [25, 26]. Besides retention prediction, QSRR also plays an important role in other fields like column characterisation [20, 21], the interpretation of retention [4, 26], and column selection (equivalent or orthogonal) [27, 28]. Given the strong performance of retention prediction, QSRR combined with high resolution mass spectrometry (HRMS) has been considered a powerful tool in metabolite analysis [4].

Some commercial software packages such as Chromsword (Merck KGaA, Darmstadt, Germany) [22, 26] or ACD/ChromGenius (ACD/Labs, Toronto, Canada), provide a limited number of molecular descriptors that are used to build QSRR models. As an alternative, a large pool of molecular descriptors can be calculated which provides more options for the construction of QSRR models [20, 22, 23]. By employing more molecular descriptors into the models, a risk of over-fitting needs to be addressed considering the fact that too many descriptors can introduce noise to the models [26, 29]. As a solution, an appropriate regression method in combination with a variable selection procedure can be used during the QSRR process, allowing only the most informative molecular descriptors to be used. Partial Least Squares (PLS), is particularly useful in handling datasets with a high number of variables compared to the number of objects, and in the presence of co-linear, redundant, and noisy variables [20, 26, 29]. In addition, as reported, the use of a Genetic Algorithm (GA) as a

variable method has shown superiority in terms of prediction accuracy and robustness of constructed QSRR models compared to other optimisation algorithms [24, 29]. Therefore, the performance of QSRR models in terms of predictive capacity can be improved significantly by employing a suitable variable selection method [29].

In Chapters 3-4 and in previous QSRR studies by the Haddad group, a variety of models have been built for retention prediction by a linear modelling technique, using partial least squares regression combined with a genetic algorithm as the variable selection method [20, 22, 23, 26]. In this chapter, retention index (RI) prediction QSRR models are examined that offer useful predictive ability for compounds having the same molecular weight, allowing false positives to be removed during the interpretation of structure identification in NTM. A novel dual-filtering approach combining structural similarity and chromatographic similarity has been employed to build suitable training sets for target analytes for the accurate prediction of RI. The elimination of false positives from the list of potential candidate compounds produced by exact mass database searches is demonstrated using the proposed QSRR approach.

5.2 Materials and methods

5.2.1 Datasets

A database originally used by Hall *et al*. to derive artificial neural network QSRR models, consisting of 1882 compounds for which retention index information and mass values are reported, was used in this study (more detail of the 1882 compounds can be found in Appendix 1, Chapter 2) [8]. In his work, they defined the RI value at the retention time of each nitro-n-alkane as 100 times the number of carbon atoms in the respective nitro-n-alkane reference compound [8]. For the original data collection, compound mixtures were analysed by HPLC/MS, where an Agilent 1100 HPLC (Agilent, Santa Clara, CA, USA) system interfaced to a QTOF-2 mass spectrometer (Waters Associates, Beverly, MA, USA) were used [8]. A Zorbax SBC18 (1 mm × 150 mm, 3.5 um particle size) column combined with a Zorbax Stable Bond (1 mm × 17 mm, 5 mm particle size) OptiGuard precolumn were used for the separation [8]. Finally, retention index (RI) values of compounds in the database were determined experimentally and the values of RI for the database compounds exhibited a wide range from 204 to 1041.

The Tanimoto similarity analysis of this database indicates a large diversity in the 1882 compounds although some isomers are found. In order to evaluate the performance of the established QSRR models in eliminating false positives having the same mass as the target metabolite, groups of representative compounds with the same mass values were selected from

the Hall *et al*. database. In all, 34 groups of compounds (see Table 2.8 in section 2.1.2, Chapter 2) were chosen, with at least five compounds having the same mass value in each group.

5.2.2 Calculation of the molecular descriptors

Molecular descriptors were calculated from the canonical simplified molecular-input line entry system strings (SMILES) of compounds using VolSurf+ 1.0.7.1 software (Molecular Discovery Ltd., Hertfordshire, UK) at a user-defined pH (pH=2.5). The process of the calculation has been detailed in section 2.3.2, Chapter 2. In all, 128 molecular descriptors were calculated following the 3D structure conversion and conformational analysis of structures. All descriptors were auto-scaled (i.e., to have zero mean and unit variance) before use for modelling process.

5.2.3 Dual-filtering

Local modelling was used throughout this study, wherein the training set of compounds used for QSRR modelling, and the QSRR model itself, were unique for each target compound for which retention index was predicted. The training set of compounds was identified using a modification of a published procedure [22, 30] where an initial subset of compounds was identified using a Tanimoto Similarity Index filter, followed by a secondary filter based on RI. For each target compound, a Tanimoto similarity analysis of the 1882 compounds in the Hall *et al*. database was carried out first using JChem for Excel (ChemAxon, Budapest, Hungary), and filtering was performed by sorting the database compounds based on their pairwise TS index in relation to the target, with the compounds having pairwise TS indices of at least 0.5 being selected as an initial training subset [26]. A correlation analysis between molecular descriptors and RI values was then performed for compounds in this initial training subset to find the descriptor which was most correlated to RI (correlation coefficient > 0.8). The initial training subset of compounds was then ranked according to the absolute difference in the value of the chosen descriptor between the initial dataset compounds and the target compound [30]. The initial training dataset compound having the lowest difference was identified as the nearest neighbour compound to the target analyte. The RI-ratio between the nearest neighbour compound and each initial training dataset compound was then calculated and a filter applied by retaining in the final training dataset only those compounds having a RI-ratio of $\geq$1.1 to the nearest neighbour. This dual-filtering process, which is illustrated schematically in Figure 5.1, ensured that the training set used in the QSRR modelling consisted only of those compounds which are structurally and chromatographically similar to the target analyte. This has been shown to greatly improve the accuracy of QSRR predictions [30].

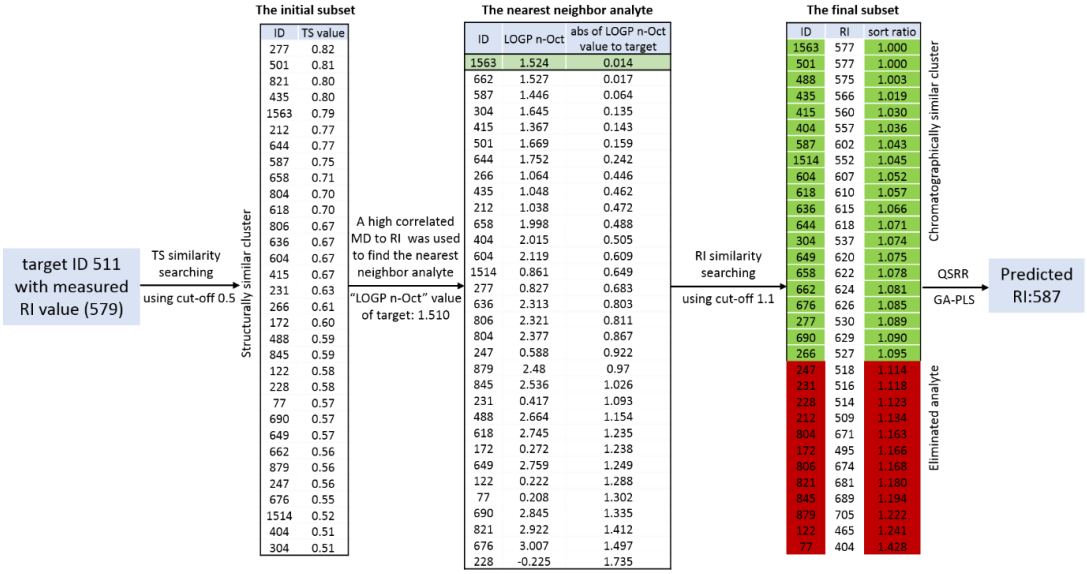


Figure 5.1. Schematic diagram of the dual-filtering process used in the present study.

5.2.4 QSRR modelling

The QSRR models predicting retention index as a function of selected molecular descriptors were obtained via partial least squares regression in combination with a genetic algorithm as the variable selection method. A popular version of a GA-PLS algorithm was used for the QSRR modelling process including descriptor selection and retention index prediction, as described in section 2.3, Chapter 2.

5.2.5 Statistics

The coefficient of determination (R^2) between the predicted and the experimental RI was calculated by constructing the corresponding scatter plot and performing a linear regression in Excel. The slope of the regression with no forced intercept was also generated with the requirement to be within the range of 0.85 to 1.15 [22, 26]. The percentage root-mean-square error of prediction (RMSEP%) of retention index for the test analyte was measured to externally validate the accuracy of GA-PLS models generated from the training set (see section 2.3.6, Chapter 2).

5.3 Results and Discussion

5.3.1 Prediction of retention index using a dual-filtering approach

Chapters 3 – 5, and previous work from the Haddad group have shown the superiority of localised models compared to global models for improving the accuracy of QSRR models [21-23, 26]. Here, a local model refers to the fact that for each compound studied, a unique QSRR model is derived. Thus, in this study, localised QSRR models were built for the 248 representative compounds, and subsequently their RI values were predicted. Specifically, each

representative compound was used successively as a test analyte by removing it from the dataset and then a training set was formed for model construction using the dual filter described under Materials and Methods [22, 26].

Tanimoto similarity (TS) searching as a filter to yield training sets has been used successfully for retention prediction in reversed-phase liquid chromatography (RPLC), hydrophilic interaction liquid chromatography (HILIC) and ion chromatography (IC) [21-23, 26]. Previous studies suggest that a training set comprising database compounds having a TS threshold of 0.5 (compared to the target analyte for which the QSRR model is to be derived) is the minimum necessary for a reliable prediction in TS-based QSRR modelling, with a higher TS similarity threshold leading to even greater accuracy of retention prediction [20, 22]. The accuracy of retention prediction for TS-based models can meet the requirement to speed up the process of chromatographic method development [20, 26], but for metabolomics studies, a more accurate prediction is highly desirable so that more false positives can be eliminated [1, 4].

In this thesis, and in previous studies by the Haddad group, in order to improve the performance of QSRR models, the role of chromatographic similarity has been investigated to define the optimal training set of compounds using as a filter the ratio between the retention time of the target compound and database compounds, with excellent prediction accuracy being obtained [21, 22, 31]. However, retention filtering is not applicable practically as the retention time for the target analyte must be known. For this reason, it becomes necessary to identify a "nearest neighbour" database compound which shows the most similar retention behaviour to the target compound in order to perform retention filtering. This approach has proved to be successful in HILIC, where a dual-filtering approach which combines the concept of structural similarity and chromatographic similarity was created for retention time. The developed dual-filtering-based QSRR models improved the retention predictability significantly with an average root mean square error in prediction (RMSEP) of 11.01% being observed [30]. Therefore, in the present study, a similar dual-filtering strategy which combined TS searching as the primary filter and RI similarity searching as the secondary filter was employed to yield suitable training sets for the generation of localised QSRR models.

The proposed dual-filtering strategy involves TS searching of the full Hall *et al*. database first, followed by RI similarity searching of the initial training dataset identified by Tanimoto similarity using the nearest neighbour in the initial training subset instead of the target analyte itself, to construct the final training dataset. The QSRR models were then derived using the dual-filtered training set for the consequent prediction of the RI of the target compound. A comparison between the measured and predicted RI values using the dual-filtering-based GA-

PLS models with the corresponding %RMSEP is depicted in Figure 5.2. In the present study, 248 compounds forming 34 representative groups sorted by molecular weight were selected for modelling and 191 of those compounds were able to be modelled using the dual-filtering approach. For the remaining 57 compounds training sets were not able to be formed as five similar compounds to the surrogate compound were not found.

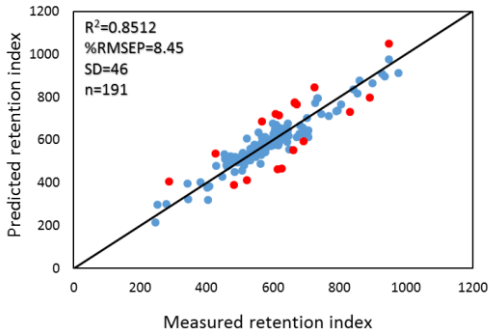


Figure 5.2. Predictive ability of dual-filter-based QSRR models for 191 analytes. For 91% of compounds (173/191) the experimental RI falls within the prediction filter (highlighted as blue). For 9% of predictions (18/191) the experimental RI falls outside the prediction filter (highlighted as red).

As shown in Figure 5.2, the predicted RI values generated using the dual-filter-based QSRR models were very well correlated with the measured data, presenting a %RMSEP value of 8.45% (which corresponds to a MAE of 33 and a standard deviation of 46). The distribution of the prediction errors of the 191 compounds is shown in Figure 5.3a and can be seen to be normally distributed. RI values of 89 compounds (46.6%) were predicted with an absolute prediction error less than 20, while 158 compounds (82.7%) were modelled with absolute prediction errors of less than 60. In addition, a residual plot of QSRR models using the dual-filtering approach has been created, showing that the linearity of the equation is supported by the uniform distributions of residuals (Figure 5.3b), where the residuals are on the vertical axis and the independent variable (measured retention index values) is on the horizontal axis. The distribution of the residuals on both sides of the zero line indicates that there is no systematic error in the obtained QSRR models.

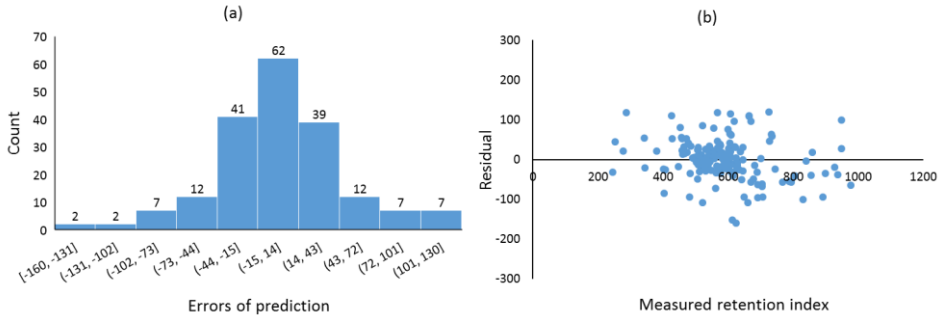


Figure 5.3. The distribution of prediction errors (a) and residual plot of prediction errors (b) using dual-filter-based GA-PLS models for 191 test analytes.

5.3.2 Retention index prediction filter

Given the strong overall performance of the constructed QSRR models using the dual-filtering approach, an evaluation was undertaken of the utility of the proposed QSRR protocol for the identification of compounds that are encountered during the interpretation process in non-targeted metabolomics. Here, RI provides information orthogonal to mass, making it possible to identify compounds from the list of candidates that have the same exact mass as the target metabolite based on the predicted RI, provided this prediction is sufficiently accurate. To achieve this, an appropriate filter range of RI for eliminating false positive identification is needed. Given that the population of errors of prediction is normally distributed (see Figure 5.3a), therefore 95% of the errors will fall within the range of ± 2 standard deviations from the mean error. The mean error and the standard deviation of the population of prediction errors were calculated, and a window of RI was then placed around each predicted retention index value (predicted RI ± 2 SD) as a filter.

Of the RI predictions, 91% were found to fall within the ± 2 SD filter (173 out of 191 compounds), 9% of predicted retention indices fell out of the filter range (18 out of 191 compounds) and are highlighted in red (see Figure 5.2). Scattered data points out of the filter range were poorly predicted with an average MAE of 111, the worse prediction of those data points may be due to the insufficient similarity of compounds in each training set to the target analyte, as the average TS index of those training sets was just slightly greater than 0.5, which was the minimum threshold of structural similarity in the dual-filtering approach.

5.3.3 Elimination of false positives

Database searching based on the exact mass is essential during the structure identification process in NTM. One of the dilemmas encountered in mass-based searching is that it may return false positives for metabolites with the same mass as the target metabolite but different behaviour in biological (and chromatographic) systems. Such false positives can be removed using the predicted retention filter, therefore narrowing down the number of candidates and accelerating the process of identification. In the present study, compounds with identical mass values were chosen to demonstrate the elimination of false positives using the ± 2 SD filter range. For each candidate molecule suggested from the mass database search, RI was predicted using the proposed QSRR approach and the predicted RI was compared with that of the target metabolite using the ± 2 SD filter range. Candidate molecules falling outside the RI filter range were classified as false positives on the basis of their RI values.

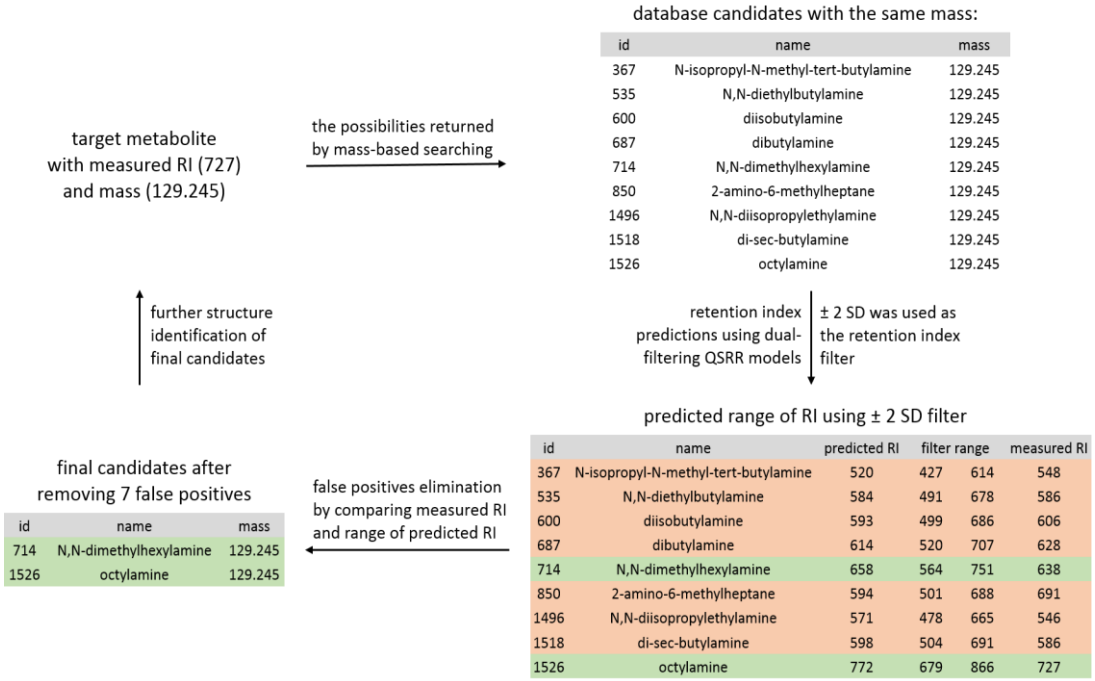


Figure 5.4. An example of the application of the retention index filter to eliminate false positives.

Figure 5.4 depicts the process of elimination of false positives using as an example a target metabolite with an observed RI of 727 and a mass of 129.245 au. Possible compounds from the database having the molecular weight of 129.245 au are listed in Figure 5.4. The values of RI for all of these possibilities were predicted using the dual-filtering approach, then a filter range of the predicted retention index ± 2 SD was applied to remove false positives. Compound N-isopropyl-N-methyl-tert-butylamine (ID 367) has a predicted RI of 520 and a filter range of 427 to 614, so we know that this is not a possibility for the target metabolite as the measured RI of the target (727) is outside that range. Similarly, in total, seven false positives were removed while only two viable candidates were retained. This filter range has been applied to the modelled compounds in each exact-mass group (34 groups in total), and 53% of groups (18 out of 34) were found where at least one false positive could be eliminated. The elimination of false positives for the representative compounds in each group is shown in Table 5.1, and more details can be found in Table 5.2. The presented method shows promise for using retention predicted by QSRR models for compound identification in non-targeted metabolomics since it allows the elimination of false positives and therefore minimises the number of candidates corresponding to the target analyte.

Table 5.1. False positives elimination using retention index filter for representative compounds in each exact-mass group

MW	Target RI	No. candidates	No. false positives eliminated
103.121	344	3	1
108.143	518	4	1
113.202	509	5	1
129.245	727	9	7
133.193	706	7	4
135.208	705	15	7
151.165	599	7	3
151.208	735	6	4
163.219	608	3	1
165.191	767	6	4
175.187	805	4	2
177.202	603	4	1
179.218	794	2	1
187.241	840	3	2
189.257	977	2	1
191.229	701	3	2
215.251	948	4	1
226.277	928	2	1

Table 5.2. Elimination of false positives for representative compounds from 18 groups

MW	ID	Measured RI	Filter range	
			Min	Max
103.121	1608	246	120.3	307.5
	1636	278	205.5	392.7
	55	344	229.7	416.9
108.143	77	404	225.6	412.8
	122	465	403.7	590.9
	218	511	413.9	601.1
	247	518	436.8	624.1
113.202	211	509	441.3	628.6
	425	563	435.9	623.1
	449	568	503.8	691
	554	593	508.7	695.9
	614	609	576.3	763.5
129.245	1496	546	477.7	664.9
	367	548	426.8	614.1
	535	586	490.6	677.8
	1518	586	504.2	691.4
	600	606	499.2	686.4
	687	628	520	707.2
	714	638	564.2	751.4
	1526	727	678.7	865.9
133.193	483	575	501.2	688.4
	520	582	491	678.2
	530	586	495.5	682.7
	550	592	493.8	681
	623	612	548.2	735.5
	652	621	559.4	746.6
	882	706	518.7	706
135.208	473	573	511.3	698.6
	488	575	518.9	706.2
	523	583	505.5	692.7
	571	598	506.9	694.1
	584	601	517.6	704.8
	597	605	496.7	683.9
	601	606	553.3	740.5
	618	610	521.9	709.1
	1521	612	506	693.2
	676	626	512.5	699.8
	690	629	527.2	714.4
	804	671	520	707.2
	821	681	572.6	759.8
	845	689	562.7	749.9
	879	705	550.6	737.8
151.165	1648	458	416	603.2
	112	461	381.1	568.3

	126	467	388.5	575.7
	1679	494	403.2	590.4
	308	538	440.9	628.1
	424	563	489.3	676.5
	1689	599	581.4	768.6
151.208	1664	560	482.6	669.8
	496	577	503.4	690.6
	598	605	576.4	763.7
	640	617	491	678.2
	737	646	524.3	711.5
	951	735	700.3	887.5
163.219	417	562	395.5	582.7
	1601	601	534.4	721.6
	609	608	484.1	671.3
165.191	104	453	439.2	626.4
	1656	521	424.1	611.3
	432	565	439	626.2
	498	577	448.5	635.7
	1691	745	627.7	814.9
	1718	767	616.3	803.5
175.187	1834	253	204.3	391.5
	1697	678	535	722.2
	945	732	701.7	888.9
	1085	805	670.8	858.1
177.202	210	509	402.5	589.7
	1796	530	439	626.2
	437	567	518.8	706
	590	603	528.6	715.8
179.218	341	544	444.7	631.9
	1693	794	643.4	830.6
187.241	849	691	534.1	721.3
	878	704	542	729.2
	1147	840	742.9	930.1
189.257	1191	859	783.5	970.7
	1396	977	818.8	1006
191.229	192	502	401.4	588.6
	572	598	480.7	667.9
	874	701	609.5	796.7
215.251	1176	854	721.2	908.4
	1260	899	770.9	958.1
	1329	936	803.9	991.1
	1351	948	881.9	1069.1
226.277	1053	789	640	827.2
	1317	928	815	1002.2

In many cases the exact-mass groups contained significant numbers of structural isomers and the proposed QSRR approach showed limited ability to exclude isomers as false positives

as these isomers generally also exhibited similar RI values. For example, 3-aminopyridine, 4-aminopyridine and 2-aminopyridine all have a mass of 94.1158 au and their RI values were 447, 456 and 484, respectively. By virtue of their structural and chromatographic similarities, such isomers will always be included in the training set. With a ± 2 SD retention index filter, no false positives could be removed in these cases.

5.3.4 Interpretation of selected descriptors

Starting from 3D maps of interaction energies between the molecule and chemical probes, VolSurf+ explores the physico-chemical property space of a molecule [32, 33]. One distinctive feature of VolSurf+ descriptors is their versatility and ease of interpretation. VolSurf+ compresses the information present in 3D maps into numerical descriptors, thus the calculation is based mainly on the generic physico-chemical properties of molecules [32, 34]. Another advantage of VolSurf+ over other software packages for descriptor generation is that all the prerequisite steps before descriptor calculation, including molecule superposition, 3D structure generation and conformational analysis are performed automatically, making descriptor calculation both user- and platform-independent and therefore less prone to error [32, 35, 36]. In addition, the number of calculated descriptors is much less than other commercial tools, so a variable selection method like GA-PLS can be performed simply and rapidly [32, 35]. VolSurf+ generates a total of 128 descriptors using different probes including OH2 probe, DRY probe, probe O and probe N1, plus non-MIF derived descriptors [32, 34, 35]. Specifically, a number of blocks of molecular descriptors were calculated and interpreted including size and shape descriptors, hydrophobic or hydrophilic regions descriptors, H-bond donor/acceptor regions descriptors, charge state and mixed descriptors, as well as 3D pharmacophoric descriptors and ADME model descriptors [32, 37-39].

A partial least squares regression combined with a genetic algorithm as the variable selection method was employed in the dual-filtering QSRR modelling. GA, as an optimisation technique inspired by the process of natural selection, has been used commonly to generate high-quality solutions to optimisation and search problems by relying on bio-inspired operators including mutation, crossover, and selection [40-42]. To avoid losing the most relevant and informative descriptors in the present study, the GA was run 100 times from different initial populations. After that, the frequency with which each descriptor was selected in the top chromosome in each run was calculated, in order to determine the best subset of descriptors to be used to build the final model. Finally, the entire process was repeated five times and the prediction errors obtained from the generated QSRR models were averaged.

The presented QSRR analysis yields models incorporating a wide range of VolSurf+ descriptors selected by the GA. Since a localised model was built for each target analyte, the

frequency of selection and the type of molecular descriptors incorporated in each QSRR model vary between target analytes. However, a brief overview of the most frequently selected descriptors appearing in the QSRR models is summarised in Table 5.3. The VolSurf+ descriptors selected in this work belong mainly to three different blocks: mixed descriptors, ADME model descriptors, and descriptors of hydrophobic regions. Not surprisingly for modelling based on reversed-phase HPLC data, Log D and Log P descriptors belonging to the family of mixed descriptors contribute most to the final QSRR models. Other significant descriptors in this family with frequent contribution to the models include the ratio between the polar surface area and the surface (PSAR), the hydrophobic surface area (HAS) and the average molecular polarizability (POL). The descriptors encoding information on the hydrophobic regions are mainly hydrophobic volumes-related. It should be also pointed out that the ADME descriptors reflecting the solubility at various pH values (LgS3-LgS11), the log Blood-Brain barrier distribution (LgBB), the intrinsic solubility (SOLY), the solubility profiling coefficients (L2LgS) and the % of protein binding (PB) were also included.

Table 5.3. Symbols and description for VolSurf+ descriptors retained most frequently in final QSRR models

Block	Symbol	Definition
Mixed descriptors	LgD9, LgD7, LgD6, LgD5, LgD10	LogD octanol/water
	LOGP n-Oct	LogP octanol/water
	LOGP c-Hex	LogP cyclohexane/water
	POL	The average molecular polarizability
	HAS	Hydrophobic Surface Area
	PSAR	The ratio between the PSA and the S
Descriptors of hydrophobic regions	D1, D3, D6, D8	Hydrophobic volumes
	CD4	Capacity Factor
ADME model descriptors	LgS11, LgS9, LgS8, LgS7.5, LgS6, LgS4, LgS3	Solubility at various pH
	LgBB	Log Blood-Brain Barrier distribution
	L2LgS	Solubility profiling coefficients
	SOLY	Intrinsic solubility
	PB	% of protein binding

5.4 Conclusions

Liquid chromatography-mass spectrometry (LC-MS), as a powerful analytical tool, has been used widely in non-targeted metabolomics. The aim of this work was to develop appropriate QSRR models which provide sufficient predictive capability to speed up the interpretation of structure identification for non-targeted metabolomic research. Here, we

show that the QSRR modelling using dual-filtering as the strategy for the generation of training sets permits a robust and highly accurate prediction of retention index. GA, as a variable selection method to choose the most informative and important descriptors, was applied successfully in combination with PLS, to build reliable QSRR models. The obtained results demonstrate that the developed QSRR strategy is capable of eliminating false positives, thereby increasing confidence of structure identification in MS-based non-targeted metabolomics. Future work includes improving the predictive ability of QSRR models by employing other molecular descriptors and using advanced similarity filtering approaches to generate training sets. The scope of modelling can also be expanded by including more compounds of greater structural similarity in the database.

5.5 References

1. Aicheler, F., J. Li, M. Hoene, R. Lehmann, G. Xu, and O. Kohlbacher, *Retention time prediction improves identification in nontargeted lipidomics approaches.* Analytical Chemistry, 2015. **87**(15): p. 7698-7704.
2. Griffiths, W.J., T. Koal, Y. Wang, M. Kohl, D.P. Enot, and H.P. Deigner, *Targeted metabolomics for biomarker discovery.* Angewandte Chemie International Edition, 2010. **49**(32): p. 5426-5445.
3. Naz, S., M. Vallejo, A. García, and C. Barbas, *Method validation strategies involved in non-targeted metabolomics.* Journal of Chromatography A, 2014. **1353**: p. 99-105.
4. Goryński, K., B. Bojko, A. Nowaczyk, A. Buciński, J. Pawliszyn, and R. Kaliszan, *Quantitative structure–retention relationships models for prediction of high performance liquid chromatography retention time of small molecules: endogenous metabolites and banned compounds.* Analytica Chimica Acta, 2013. **797**: p. 13-19.
5. Creek, D.J., W.B. Dunn, O. Fiehn, J.L. Griffin, R.D. Hall, Z. Lei, R. Mistrik, S. Neumann, E.L. Schymanski, and L.W. Sumner, *Metabolite identification: are you sure? And how do your peers gauge your confidence?* Metabolomics, 2014. **10**(3): p. 350-353.
6. Brown, M., W.B. Dunn, P. Dobson, Y. Patel, C. Winder, S. Francis-McIntyre, P. Begley, K. Carroll, D. Broadhurst, and A. Tseng, *Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics.* Analyst, 2009. **134**(7): p. 1322-1332.
7. Fuhrer, T. and N. Zamboni, *High-throughput discovery metabolomics.* Current Opinion in Biotechnology, 2015. **31**: p. 73-78.
8. Hall, L.M., D.W. Hill, L.C. Menikarachchi, M.-H. Chen, L.H. Hall, and D.F. Grant, *Optimizing artificial neural network models for metabolomics and systems biology: an example using HPLC retention index data.* Bioanalysis, 2015. **7**(8): p. 939-955.
9. Díaz, R., H. Gallart-Ayala, J.V. Sancho, O. Nunez, T. Zamora, C.P. Martins, F. Hernández, S. Hernández-Cassou, J. Saurina, and A. Checa, *Told through the wine: A liquid chromatography–mass spectrometry interplatform comparison reveals the influence of the global approach on the final annotated metabolites in non-targeted metabolomics.* Journal of Chromatography A, 2016. **1433**: p. 90-97.
10. Meyer, M.R. and H.H. Maurer, *Current applications of high-resolution mass spectrometry in drug metabolism studies.* Analytical and Bioanalytical Chemistry, 2012. **403**(5): p. 1221-1231.
11. Virus, E., T. Sobolevsky, and G. Rodchenkov, *Introduction of HPLC/orbitrap mass spectrometry as screening method for doping control.* Journal of Mass Spectrometry, 2008. **43**(7): p. 949-957.

12. García-Lavandeira, J., B. Losada, J. Martínez-Pontevedra, M. Lores, and R. Cela, *Computer-assisted method development in liquid chromatography–mass spectrometry: New proposals.* Journal of Chromatography A, 2008. **1208**(1): p. 116-125.
13. Mairinger, T., T.J. Causon, and S. Hann, *The potential of ion mobility–mass spectrometry for non-targeted metabolomics.* Current Opinion in Chemical Biology, 2018. **42**: p. 9-15.
14. Bueno, M.J.M., F.J. Díaz-Galiano, Ł. Rajski, V. Cutillas, and A.R. Fernández-Alba, *A non-targeted metabolomic approach to identify food markers to support discrimination between organic and conventional tomato crops.* Journal of Chromatography A, 2018.
15. Ganna, A., S. Salihovic, J. Sundström, C.D. Broeckling, Å.K. Hedman, P.K. Magnusson, N.L. Pedersen, A. Larsson, A. Siegbahn, and M. Zilmer, *Large-scale metabolomic profiling identifies novel biomarkers for incident coronary heart disease.* PLoS Genetics, 2014. **10**(12): p. e1004801.
16. Christians, U., J. Klawitter, A. Hornberger, and J. Klawitter, *How unbiased is non-targeted metabolomics and is targeted pathway screening the solution?* Current Pharmaceutical Biotechnology, 2011. **12**(7): p. 1053-1066.
17. Putri, S.P., S. Yamamoto, H. Tsugawa, and E. Fukusaki, *Current metabolomics: technological advances.* Journal of Bioscience and Bioengineering, 2013. **116**(1): p. 9-16.
18. Ghasemi, J. and S. Saaidpour, *QSRR prediction of the chromatographic retention behavior of painkiller drugs.* Journal of Chromatographic Science, 2009. **47**(2): p. 156-163.
19. Schefzick, S., C. Kibbey, and M.P. Bradley, *Prediction of HPLC conditions using QSPR techniques: an effective tool to improve combinatorial library design.* Journal of Combinatorial Chemistry, 2004. **6**(6): p. 916-927.
20. Talebi, M., S.H. Park, M. Taraji, Y. Wen, R.I. Amos, P.R. Haddad, R. Shellie, R. Szucs, C. Pohl, and J.W. Dolan, *Retention time prediction based on molecular structure in pharmaceutical method development: A perspective.* LCGC North America, 2016. **34**(8): p. 550-558.
21. Taraji, M., P.R. Haddad, R.I. Amos, M. Talebi, R. Szucs, J.W. Dolan, and C.A. Pohl, *Prediction of retention in hydrophilic interaction liquid chromatography using solute molecular descriptors based on chemical structures.* Journal of Chromatography A, 2017. **1486**: p. 59-67.
22. Tyteca, E., M. Talebi, R. Amos, S.H. Park, M. Taraji, Y. Wen, R. Szucs, C.A. Pohl, J.W. Dolan, and P.R. Haddad, *Towards a chromatographic similarity index to establish localized quantitative structure-retention models for retention prediction: use of retention factor ratio.* Journal of Chromatography A, 2017. **1486**: p. 50-58.
23. Park, S.H., P.R. Haddad, M. Talebi, E. Tyteca, R.I. Amos, R. Szucs, J.W. Dolan, and C.A. Pohl, *Retention prediction of low molecular weight anions in ion chromatography based on quantitative structure-retention relationships applied to the linear solvent strength model.* Journal of Chromatography A, 2017. **1486**: p. 68-75.
24. Žuvela, P., J.J. Liu, K. Macur, and T. Baczek, *Molecular descriptor subset selection in theoretical peptide quantitative structure–retention relationship model development using nature-inspired optimization algorithms.* Analytical Chemistry, 2015. **87**(19): p. 9876-9883.
25. Héberger, K., *Quantitative structure–(chromatographic) retention relationships.* Journal of Chromatography A, 2007. **1158**(1-2): p. 273-305.
26. Wen, Y., M. Talebi, R.I. Amos, R. Szucs, J.W. Dolan, C.A. Pohl, and P.R. Haddad, *Retention prediction in reversed phase high performance liquid chromatography using quantitative structure-retention relationships applied to the Hydrophobic Subtraction Model.* Journal of Chromatography A, 2018. **1541**: p. 1-11.
27. Snyder, L., J. Dolan, and P. Carr, *The hydrophobic-subtraction model of reversed-phase column selectivity.* Journal of Chromatography A, 2004. **1060**(1): p. 77-116.
28. Græsbøll, R., N.J. Nielsen, and J.H. Christensen, *Using the hydrophobic subtraction model to choose orthogonal columns for online comprehensive two-dimensional liquid chromatography.* Journal of Chromatography A, 2014. **1326**: p. 39-46.

29. Talebi, M., G. Schuster, R.A. Shellie, R. Szucs, and P.R. Haddad, *Performance comparison of partial least squares-related variable selection methods for quantitative structure retention relationships modelling of retention times in reversed-phase liquid chromatography*. Journal of Chromatography A, 2015. **1424**: p. 69-76.
30. Taraji, M., P.R. Haddad, R.I. Amos, M. Talebi, R. Szucs, J.W. Dolan, and C.A. Pohl, *Use of dual-filtering to create training sets leading to improved accuracy in quantitative structure-retention relationships modelling for hydrophilic interaction liquid chromatographic systems.* Journal of Chromatography A, 2017. **1507**: p. 53-62.
31. Park, S.H., P.R. Haddad, R.I. Amos, M. Talebi, R. Szucs, C.A. Pohl, and J.W. Dolan, *Towards a chromatographic similarity index to establish localised Quantitative Structure-Retention Relationships for retention prediction. III Combination of Tanimoto similarity index, logP, and retention factor ratio to identify optimal analyte training sets for ion chromatography*. Journal of Chromatography A, 2017. **1520**: p. 107-116.
32. Cruciani, G., M. Pastor, and W. Guba, *VolSurf: a new tool for the pharmacokinetic optimization of lead compounds.* European Journal of Pharmaceutical Sciences, 2000. **11**: p. S29-S39.
33. Cramer, R.D., D.E. Patterson, and J.D. Bunce, *Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins.* Journal of the American Chemical Society, 1988. **110**(18): p. 5959-5967.
34. Clementi, S., G. Cruciani, P. Fifi, D. Riganelli, R. Valigi, and G. Musumarra, *A new set of principal properties for heteroaromatics obtained by GRID.* Molecular Informatics, 1996. **15**(2): p. 108-120.
35. Cruciani, G., M. Pastor, and S. Clementi, *Handling information from 3D grid maps for QSAR studies*, in *Molecular modeling and prediction of bioactivity*. 2000, Springer. p. 73-81.
36. Crivori, P., G. Cruciani, P.-A. Carrupt, and B. Testa, *Predicting blood− brain barrier permeation from three-dimensional molecular structure.* Journal of Medicinal Chemistry, 2000. **43**(11): p. 2204-2216.
37. Guba, W. and G. Cruciani, *Molecular field-derived descriptors for the multivariate modeling of pharmacokinetic data*, in *Molecular modeling and prediction of bioactivity*. 2000, Springer. p. 89-94.
38. Cruciani, G., P. Crivori, P.-A. Carrupt, and B. Testa, *Molecular fields in quantitative structure–permeation relationships: the VolSurf approach.* Journal of Molecular Structure: THEOCHEM, 2000. **503**(1): p. 17-30.
39. Crivori, P., G. Cruciani, P.-A. Carrupt, and B. Testa, *Predicting blood− brain barrier permeation from three-dimensional molecular structure.* Journal of Medicinal Chemistry, 2000. **43**(11): p. 2204-2216.
40. Leardi, R. and A.L. Gonzalez, *Genetic algorithms applied to feature selection in PLS regression: how and when to use them.* Chemometrics and Intelligent Laboratory Systems , 1998. **41**(2): p. 195-207.
41. Vainio, M.J. and M.S. Johnson, *Generating conformer ensembles using a multiobjective genetic algorithm.* Journal of Chemical Information and Modeling, 2007. **47**(6): p. 2462-2474.
42. Leardi, R., *Application of genetic algorithm-PLS for feature selection in spectral data sets.* Journal of Chemometrics, 2000. **14**(5-6): p. 643-655.

6 General Conclusions

Traditional method development in the pharmaceutical industry is mainly carried out by trial-and-error laboratory experimentation, which is time-consuming and costly [1-3]. Moreover, given the fact that a wide variety of equipment, columns, eluent and operational parameters are involved, this means that the developed chromatographic methods may not be inherently robust and can be poorly understood [4, 5]. Within the broad classification of liquid chromatography, reversed-phase liquid chromatography (RPLC) is still the most widely used separation technique in the pharmaceutical and other industries, but other chromatographic techniques including hydrophilic-interaction chromatography (HILIC) or ion chromatography (IC) are also frequently used as complementary methods [6-8]. The detailed retention mechanism applicable in RPLC has been given intensive study which is useful to the process of method development [8, 9], but considering the excessive number of stationary phases available for RPLC, method development in RPLC is still a complex task [8, 10, 11]. Computer-assisted approaches can accelerate method development by accurately predicting the retention of analytes, leading to optimisation of chromatographic performance [12]. There is a strong demand for a tool to predict chromatographic behaviour, and as a result, to speed up the scoping phase of chromatographic method development [5, 12]. This objective can be achieved using a combination of analytical routines and chemometric techniques, and quantitative structure-retention relationship (QSRR) modelling, which has been has already found some use for retention prediction, is a promising solution from a variety of chemometric methods [13, 14]. QSRR methodology aims to find the mathematical relationships between the chromatographic parameters and the molecular descriptors that are generated based on the structures of the analytes [12, 15, 16].

This thesis describes the development of retention prediction models using QSRR methodology for a variety of differently structured pharmaceutical compounds and commercially available stationary phases employed in the RPLC mode. In addition, the application of the proposed QSRR approach in the prediction of co-elution, retention prediction for new compounds, and the elimination of the false positives in Non-Targeted Metabolomics (NTM) was explored and investigated.

Firstly, the QSRR models were developed to directly predict the retention times of compounds in three datasets obtained from the literature and an internet open database on different RPLC stationary phases, with a view to selecting the most suitable stationary phase(s) for the separation of a given set of compounds. The study was conducted using compounds in three databases as target compounds. Molecular descriptors of each compound were generated using Dragon 6.0 software (Talete, Milano, Italy), based on their chemical structures that had

been optimised using density functional theory. Given the huge number of molecular descriptors calculated, a genetic algorithm (GA) was employed to choose the most informative molecular descriptors for the subsequent QSRR modelling process. Those selected molecular descriptors were then used to form QSRR models *via* a partial least squares (PLS) regression. The constructed models were then used to predict the retention times for the target compounds.

For this part of the study, several strategies for compound filtering to identify the optimal training set of compounds for QSRR calculations, including the ratio of retention factor (*k*-ratio), the Tanimoto similarity (TS), the log D and log P similarity, plus a dual-filtering strategy, were evaluated and compared. Local QSRR models generated using the above filtering approaches were built to predict retention. Among the constructed QSRR models, the filter that used the ratio of retention factor appeared to be the most effective approach to minimising prediction errors. However, the *k*-ratio filter was impractical as it required the retention of the target compound to be known before modelling, therefore it cannot be used directly in QSRR modelling for retention prediction. However, the *k*-ratio filter is relevant as a benchmark for the minimum achievable prediction errors in the QSRR modelling, and to compare the performance of other more practical filters. For the QSRR models using Tanimoto similarity as the filter, training sets using the top ten most similar compounds resulted in much higher prediction errors than those using the *k*-ratio similarity filter. The prediction errors yielded using the log D filter, which represents the hydrophobic interaction in reversed-phase retention, were slightly improved compared to the Tanimoto filter, but the log P filter did not produce acceptably low prediction errors. The dual-filtering method, in which Tanimoto similarity (or, alternatively, log D) was taken as the primary filter, then combined with *k*-ratio similarity as the secondary filter, did not lead to meaningful improvements in retention prediction. It is worth noting that the low average similarity score of the top ten compounds used in the training sets contributed most to the poor retention prediction when the Tanimoto filter was applied. The low average similarity score also indicated that a larger and more homogenous database was highly desirable for the construction of training sets, allowing sufficient numbers of compounds with a high pair-wise similarity to be found when using the Tanimoto filter. The dual filter was still based on the similarity of chemical structures of compounds, and therefore its accuracy could also be improved in the case of larger and more homogenous datasets by including more compounds with greater chromatographic similarity.

Next, retention prediction in RPLC was performed indirectly using the Hydrophobic Subtraction Model (HSM). QSRR models were developed to predict solute coefficients of the HSM using several compound classification approaches and employing different resources of

molecular descriptors. Both global model and local models were built for the prediction. Filtering approaches to yield local models including the Leave-One-Out (LOO) approach, the compound type (LCT) approach, and the secondary dominant interaction after hydrophobicity (LSDI) approach. In this part of the work, molecular descriptors of compounds were calculated using both Dragon (Talete, Milano, Italy), and VolSurf+ (Molecular Discovery Ltd., Hertfordshire, UK). Finally, the five predicted solute coefficients combined with the five known column parameters were fitted into the HSM to calculate the retention of the given target compounds. Among the above-mentioned filtering approaches, the LSDI showed the best prediction for the solute coefficients compared to other filters, followed by the LCT. The LOO and the global approaches resulted in poor prediction for solute coefficients. It can be expected that better predictions of solute coefficients would be obtained using the LCT and the LSDI approach because compound classification was involved which allowed better suited training sets to be constructed. In terms of the comparison of the resources of molecular descriptors, no significant difference was observed between Dragon, VolSurf+, and combined descriptors. Since the number of molecular descriptors generated from VolSurf+ was much less than for Dragon (VolSurf+ generated 128 descriptors only), therefore the genetic algorithm (GA) as a variable selection method can be performed more rapidly, making VolSurf+ a more attractive option for the subsequent modelling. Additionally, the importance of the hydrophobic interaction term of the HSM to retention prediction was investigated. Results showed that an approximate HSM using the hydrophobicity term only yielded comparable results to those using the full HSM which consisted of five terms.

To further investigate the feasibility of using the HSM and QSRR for retention prediction in RPLC, a combined retention dataset of 148 compounds was employed. Retention prediction was achieved through modelling only the solute coefficient η' representing solute hydrophobicity in the HSM. Molecular descriptors were calculated using VolSurf+ only. Results showed improvement in prediction accuracy for the LSDI, the LCT, and the local Tanimoto similarity (LTS) approaches compared with global models derived from the whole dataset without filtering. Of these approaches, the LCT exhibited advantages in its ease of application and the larger number of compounds which could be modelled using this approach. However, the LTS approach was the simplest to apply, and provided that there was a sufficient number of similar compounds in the dataset, it yielded sufficiently accurate results. It is recognised that modelling based on η' can predict only the hydrophobic selectivity of analytes and that the predicted elution order of target compounds was determined solely by their η' values. Nevertheless, the predictions reported here showed sufficient accuracy to meet the major objective of this study, namely to determine the likelihood of co-elution of analytes in applications such as early stage drug development.

Given the robust performance of the constructed QSRR models, the predictive ability for new compounds on columns that had never been used in the modelling process was also investigated. Five analytical grade compounds: pindolol, 1, 8-dihydroxynaphthalene, 4-ethylnitrobeneze, 2-phenylbutane and 4-heptylbenzoic acid were investigated. The five columns employed were all obtained from Thermo Fisher Scientific: an AcclaimTM 120 C18 column (4.6 × 150 mm, 5.0 μm); an Accucore C18 column (4.6 × 150 mm, 2.6 μm); a Hypersil GOLD C8 column (4.6 × 150 mm, 5.0 μm); a Hypersil GOLD C18 column (4.6 × 150 mm, 5.0 μm) and a Hypersil ODS C18 column (4.6 × 150 mm, 5.0 μm). After retention data collection of the five new compounds on the five new columns, their retention times were predicted using the proposed QSRR method *via* the approximate HSM. Very good correlations between the predicted and measured retention times of the five test compounds using the LCT and the LSDI approaches were obtained. R^2 values (between predicted and observed retention times) of 0.9634 and 0.9861 were observed for the LSDI and the LCT approach, respectively. In terms of prediction error, the LCT approach generated a %RMSEP of 15.41, the LSDI approach resulted in a %RMSEP of 28.67. The powerful performance of the QSRR models illustrated the feasibility of using the proposed approaches to predict retention of new compounds.

QSRR has been applied in many fields, including the characterisation of columns, the interpretation of retention mechanisms and retention prediction of compounds [13, 14]. Considering this advantage, a predictive QSRR methodology combined with high-resolution mass spectrometry (HRMS) could be a powerful tool in metabolite analysis [17]. For an LC-MS-based database, chromatographic retention has been used as a feature for the identification of metabolites. In MS-based non-targeted metabolomics (NTM) analysis, an accurate mass is usually used to define molecular formulae by searching electronic resources [17-19]. Based on the exact mass of the metabolite, potential candidates can be found from the database, but the limitation of mass searching is that in many cases, an excessive number of candidates are returned, meaning that false positives need to be removed [17, 20]. In the present study, QSRR models that offered robust predictive ability for compounds with the same molecular weight were developed using a novel dual-filtering approach which combined structural and chromatographic similarity, allowing false positives to be removed using predicted retention which provides information orthogonal to mass during the interpretation of structure identification in non-targeted metabolomics.

The proposed dual-filtering strategy involved Tanimoto searching first, followed by retention similarity searching using the nearest neighbour in the initial subset instead of the target analyte itself, to construct the final training set. A retention index database of 1882

compounds with known molecular weights was employed. VolSurf+ molecular descriptors were calculated for 1882 compounds, and 34 groups of compounds (248 compounds in total) were chosen as target compounds with at least five compounds having the same mass value in each group. Dual-filtering was used for each target compound to build a local training set and these constructed training sets were used to form QSRR models. Finally, 191 of those compounds were modelled using the dual-filtering approach. For the remaining 57 compounds training sets were not able to be formed, as five similar compounds to the target compound were not found. The dual-filter-based QSRR models generated good retention prediction with a %RMSEP value of 8.45%, 89 compounds (46.6%) were accurately predicted with an absolute prediction error of retention index less than 20, while 158 compounds (82.7%) were modelled with absolute prediction errors of retention index of less than 60.

To evaluate the feasibility for the elimination of false positives, a retention index window was placed around each predicted retention index (predicted RI ± 2 SD) as a filter used to eliminate false positives. Results showed that 91% of retention index predictions fell within the ± 2 SD filter (173 out of 191 compounds), with 9% of predicted retention indices falling out of the filter range (18 out of 191 compounds). In terms of elimination, a target analyte was identified first, and the retentions of other analytes with identical molecular weight were predicted. Then the retention indices were compared using the above retention index filter to eliminate those where the target analyte RI fell outside the predicted RI ± 2 SD range of the possible matches. After applying this retention index filter range to the modelled compounds in each group (34 groups in total), 53% of groups (18 out of 34) were found where at least one false positive could be eliminated. The presented method shows promise for using retention predicted by QSRR models for compound identification in non-targeted metabolomics since it allows the elimination of false positives and therefore minimises the number of candidates corresponding to the target analyte.

However, the proposed approach showed limited ability to eliminate false positives when these compounds were structural isomers of the target compound, since these isomers exhibited similar retention indices to the target compound. Possible methods to solve this issue include improving the performance of the QSRR models by employing more compounds in the database, or providing more accurate RI predictions using better similarity filters, allowing sufficient training sets to be built with an elevated level of similar compounds to the target.

This thesis comprised the development of QSRR models for retention prediction and the application of the proposed QSRR in non-targeted metabolomics. The QSRR strategy was applied successfully to RPLC with good accuracy in retention prediction and thereby showed good potential to accelerate the process of chromatographic method development.

Furthermore, with the accurate retention prediction of compounds having the same molecular weight, the proposed QSRR methodology can be seen to be a useful tool in eliminating false positives, and improving the confidence and the efficiency of metabolite identification in NTM.

References

1. Krisko, R.M., K. McLaughlin, M.J. Koenigbauer, and C.E. Lunte, *Application of a column selection system and DryLab software for high-performance liquid chromatography method development.* Journal of Chromatography A, 2006. **1122**(1): p. 186-193.
2. Bolanča, T., Š. Ukić, M. Novak, and M. Rogošić, *Computer assisted method development in liquid chromatography.* Croatica Chemica Acta, 2014. **87**(2): p. 111-122.
3. Snyder, L.R., J.J. Kirkland, and J.L. Glajch, *Practical HPLC method development*. 2012: John Wiley & Sons.
4. Tyteca, E., M. Talebi, R. Amos, S.H. Park, M. Taraji, Y. Wen, R. Szucs, C.A. Pohl, J.W. Dolan, and P.R. Haddad, *Towards a chromatographic similarity index to establish localized quantitative structure-retention models for retention prediction: use of retention factor ratio.* Journal of Chromatography A, 2017. **1486**: p. 50-58.
5. Talebi, M., S.H. Park, M. Taraji, Y. Wen, R.I. Amos, P.R. Haddad, R. Shellie, R. Szucs, C. Pohl, and J.W. Dolan, *Retention time prediction based on molecular structure in pharmaceutical method development: A perspective.* LCGC North America, 2016. **34**(8): p. 550-558.
6. Borges, E.M., *How to select equivalent and complimentary reversed phase liquid chromatography columns from column characterization databases.* Analytica Chimica Acta, 2014. **807**: p. 143-152.
7. Snyder, L., J. Dolan, and P. Carr, *The hydrophobic-subtraction model of reversed-phase column selectivity.* Journal of Chromatography A, 2004. **1060**(1): p. 77-116.
8. Dong, M.W., *A Universal Reversed-Phase HPLC Method for Pharmaceutical Analysis.* LCGC North America, 2016. **34**(6): p. 408-419.
9. Wilson, N., M. Nelson, J. Dolan, L. Snyder, R. Wolcott, and P. Carr, *Column selectivity in reversed-phase liquid chromatography: I. A general quantitative relationship.* Journal of Chromatography A, 2002. **961**(2): p. 171-193.
10. Karmarkar, S., R. Garber, Y. Genchanok, S. George, X. Yang, and R. Hammond, *Quality by design (QbD) based development of a stability indicating HPLC method for drug and impurities.* Journal of Chromatographic Science, 2011. **49**(6): p. 439-446.
11. Kormány, R., I. Molnár, J. Fekete, D. Guillarme, and S. Fekete, *Robust UHPLC separation method development for multi-API product containing amlodipine and bisoprolol: the impact of column selection.* Chromatographia, 2014. **77**(17-18): p. 1119-1127.
12. Wen, Y., M. Talebi, R.I. Amos, R. Szucs, J.W. Dolan, C.A. Pohl, and P.R. Haddad, *Retention prediction in reversed phase high performance liquid chromatography using quantitative structure-retention relationships applied to the Hydrophobic Subtraction Model.* Journal of Chromatography A, 2018. **1541**: p. 1-11.
13. Goodarzi, M., R. Jensen, and Y. Vander Heyden, *QSRR modeling for diverse drugs using different feature selection methods coupled with linear and nonlinear regressions.* Journal of Chromatography B, 2012. **910**: p. 84-94.
14. Ghasemi, J. and S. Saaidpour, *QSRR prediction of the chromatographic retention behavior of painkiller drugs.* Journal of Chromatographic Science, 2009. **47**(2): p. 156-163.
15. Taraji, M., P.R. Haddad, R.I. Amos, M. Talebi, R. Szucs, J.W. Dolan, and C.A. Pohl, *Prediction of retention in hydrophilic interaction liquid chromatography using solute*

molecular descriptors based on chemical structures. Journal of Chromatography A, 2017. **1486**: p. 59-67.
16. Park, S.H., P.R. Haddad, M. Talebi, E. Tyteca, R.I. Amos, R. Szucs, J.W. Dolan, and C.A. Pohl, *Retention prediction of low molecular weight anions in ion chromatography based on quantitative structure-retention relationships applied to the linear solvent strength model.* Journal of Chromatography A, 2017. **1486**: p. 68-75.
17. Goryński, K., B. Bojko, A. Nowaczyk, A. Buciński, J. Pawliszyn, and R. Kaliszan, *Quantitative structure–retention relationships models for prediction of high performance liquid chromatography retention time of small molecules: endogenous metabolites and banned compounds.* Analytica Chimica Acta, 2013. **797**: p. 13-19.
18. Naz, S., M. Vallejo, A. García, and C. Barbas, *Method validation strategies involved in non-targeted metabolomics.* Journal of Chromatography A, 2014. **1353**: p. 99-105.
19. Bueno, M.J.M., F.J. Díaz-Galiano, Ł. Rajski, V. Cutillas, and A.R. Fernández-Alba, *A non-targeted metabolomic approach to identify food markers to support discrimination between organic and conventional tomato crops.* Journal of Chromatography A, 2018.
20. Aicheler, F., J. Li, M. Hoene, R. Lehmann, G. Xu, and O. Kohlbacher, *Retention time prediction improves identification in nontargeted lipidomics approaches.* Analytical Chemistry, 2015. **87**(15): p. 7698-7704.

Appendix

Appendix 1. Database of 1882 compounds

1882 compounds with known molecular weight (MW) and retention index (RI) was employed for the Non-Targeted Metabolomics (NTM) analysis [1].

Appendix 1. Molecular weights (MW) and retention indices (RI) of 1882 compounds

ID	Name	MW	RI
1	tris(phosphonomethyl)amine	299.051	204
2	bis(phosphonomethyl){2-[bis(phosphonomethyl)amino]ethyl}amine	436.126	206
3	(2,3-diphosphonopropyl)diethylamine	275.178	211
4	5-amino-2,6-dioxo-1,3-dihydropyrimidine-4-sulfonic_acid	207.161	213
5	3-oxo-1,2,4-triazoline-5-carboxylic_acid	129.075	216
6	2,6-dioxo-1,3-dihydropyrimidine-4-carboxylic_acid	156.098	225
7	urea	60.055	227
8	3-phosphonopropylamine	139.091	230
9	4,5-dihydroxyimidazolidin-2-one	118.092	233
10	3-aminobenzenesulfonic_acid	173.186	238
11	2-(aminocarbonylamino)-3-carbamoylpropanoic_acid	175.144	239
12	pyrimidine-2,4-diol	112.088	243
13	1-({[(2-oxopyrrolidinyl)methyl](hydroxyphosphoryl)}methyl)pyrrolidin-2-one	260.229	245
14	1-methylpyridine-3-carboxylic_acid_chloride	138.146	245
15	N-[1-(acetylamino)-2,3,4,5,6-pentahydroxyhexyl]acetamide	280.277	246
16	5-methyl-2,6-dioxo-1,3-dihydropyrimidine-4-carboxylic_acid	170.124	247
17	1-methyl-2,6-dioxo-1,3-dihydropyrimidine-4-carboxylic_acid	170.124	248
18	2-(5-carbamoyl-2,4-dioxo-1,3-dihydropyrimidinyl)acetic_acid	213.149	249
19	2-(acetylamino)butanedioic_acid	175.141	251
20	2-oxo-4-imidazoline-4-carboxylic_acid	128.087	254
21	acetamide	59.068	260
22	N,N-dimethylformamide	73.094	276
23	N-methylformamide	59.068	276
24	2,3-dihydroxybutanedihydrazide	178.147	278
25	pyrimidin-2-ol	96.0884	278
26	1,2,4-triazole	69.0658	284
27	N-[di(acetylamino)methyl]acetamide	187.198	284
28	pyrimidine-4-carboxylic_acid	124.099	287
29	1,3-dimethylurea	88.109	290
30	2-(2,4-dioxo(1,3-dihydropyrimidinyl))-N-(2-hydroxyethyl)acetamide	213.193	290
31	5-methyl-1H-1,2,3,4-tetraazole	84.0804	292
32	2-(2,4-dioxo-1,3-dihydropyrimidin-5-yl)acetic_acid	170.124	293
33	5-oxo-1-(4-sulfophenyl)-2-pyrazoline-3-carboxylic_acid	284.243	294
34	1H-1,2,3-triazole	69.0658	296
35	3,4,5-trihydroxybenzamide	169.137	298
36	3-(2,5-dioxo-1,3-diazolidin-4-yl)propanoic_acid	172.14	301
37	4-hydroxybutanehydrazide	118.135	306
38	5,6-dimethylpyrazine-2,3-dicarboxylic_acid	196.162	314
39	2-(acetylamino)propanoic_acid	131.131	315
40	5-(trimethylamino)-3-hydropyrimidin-2-one	154.191	320
41	(2S,3S,5S,1R)-5-amino-3-(hydroxymethyl)cyclopentane-1,2-diol_chloride	147.174	321

42	3-[(methylamino)methyl]-1,2,4-triazolin-5-one	128.133	326
43	2-(methylamino)ethan-1-ol	75.11	328
44	2,3-dihydroxy-N,N-dimethyl-N',N'-dimethylbutane-1,4-diamide	204.225	328
45	5-hydro-4-imidazolino[4,5-c]pyridine-2,4-dione	151.124	329
46	pyrazino[2,3-d]pyridazine-5,8-diol	164.123	329
47	3-amino-6-methyl-4H-1,2,4-triazin-5-one	126.118	333
48	2-methoxypropanamide	103.121	334
49	6-amino-2-methyl-3-hydropyrimidin-4-one	125.13	335
50	piperazine-1-carbaldehyde	114.147	337
51	piperidin-4-ol	101.148	338
52	5-(hydroxymethyl)pyrrolidin-3-ol	117.147	340
53	1-methyl-5-oxopyrrolidine-3-carboxylic_acid	143.142	341
54	4-methylpyrimidin-2-ol	110.115	344
55	gamma aminobutyric acid	103.121	344
56	imidazole	68.078	344
57	3-aminohydropyridin-2-one	110.115	346
58	pyridine-3,5-dicarboxylic_acid	167.121	347
59	1-methyl-6-oxo-1,4,5-trihydropyridazine-3-carboxylic_acid	156.141	352
60	pyridine-2,6-dicarboxylic_acid	167.121	354
61	3-aminopiperidin-2-one	114.147	361
62	N-(4-pyridylcarbonylamino)acetamide	179.178	363
63	4-aminopyrimidine	95.1036	365
64	piperazine-1,4-dicarbaldehyde	142.157	367
65	1-methylimidazole	82.1048	378
66	amino-N-(2-pyridylmethyl)amide	151.168	379
67	4-aminobutan-2-ol	89.137	381
68	6-amino-3-(2,3-dihydroxypropyl)-3-hydropyrimidin-2-one	185.182	394
69	6-oxohydropyridine-3-carboxylic_acid	139.11	394
70	1,3,5-triazine-2,4,6-triamine	126.121	396
71	2H-3,4,5,6-tetrahydropyran-4-ylamine	101.148	396
72	3-pyridylmethan-1-ol	109.127	396
73	2-(hydroxymethyl)-2-(2-pyridyl)propane-1,3-diol	183.207	399
74	2-methyl-1-pyrroline	83.1328	399
75	2-amino-2-methylpropan-1-ol	89.137	401
76	2-methylimidazole	82.1048	404
77	p-phenylenediamine	108.143	404
78	pyridazine	80.089	405
79	(2-methoxyethyl)(methyl)amine	89.137	408
80	methylthiocarboxamidine_iodide	90.1428	409
81	propylamine	59.1108	415
82	pyridin-4-ol	95.1006	415
83	1,3-dimethyl-2,4-dioxo-1,3-dihydroquinazoline-6-sulfonic_acid	270.259	418
84	1-methyl-6-oxohydropyridine-4-carboxylic_acid	153.137	420
85	1-(2-(1,2,4-triazolyl)ethyl)-1,2,4-triazole	164.169	421
86	3-oxo-N-[2-(3-oxobutanoylamino)ethyl]butanamide	228.247	421
87	5-(3-pyridyl)-1H-1,2,3,4-tetraazole	147.139	424
88	2-(3-methyl-2,4-dioxo-1,3-dihydropyrimidin-5-yl)acetic_acid	184.151	425
89	3-amino-2-methylbenzamide	150.18	426
90	diethylamine	73.1376	426
91	1,3-thiazoline-2-ylamine	102.154	431
92	2H-3,4,5,6-tetrahydropyran-3-carboxamide	129.158	432
93	N-[(acetylamino)-4-pyridylmethyl]acetamide	207.232	434
94	piperidine-2-carboxamide	128.174	437

95	purine	120.113	438
96	4-[(carbamoylmethyl)amino]benzamide	193.205	439
97	6-amino-5-methyl-3-hydropyrimidin-2-one	125.13	441
98	2,5-diaminopyridine	109.13	442
99	2-aminopurine	135.128	446
100	2-methylquinoline-4-carboxylic_acid	187.198	447
101	3-aminopyridine	94.1158	447
102	2-[(propan-2-yl)amino]ethan-1-ol	103.164	448
103	3-amino-3-methylbutan-2-one	101.148	451
104	3-amino-3-phenylpropanoic_acid	165.191	453
105	pyridazine-3,4,5,6-tetraamine	140.147	454
106	[6-(hydroxymethyl)-2-pyridyl]methan-1-ol	139.154	455
107	4(5)-methylimidazole	82.1048	455
108	4-aminopyridine	94.1158	456
109	N-(2-methyl-6-oxohydropyrimidin-4-yl)acetamide	167.167	456
110	3-(2-hydroxyethyl)-4,6-dimethyl-3-hydropyrimidin-2-one	168.195	458
111	N,N-dimethylallylamine	85.1486	459
112	2-(4-hydroxyphenyl)acetamide	151.165	461
113	3-methoxypropan-1-amine	89.137	461
114	pyrimidine	80.089	461
115	(1,4-dioxan-2-ylmethyl)methylamine	131.174	462
116	2-(cyclohexylamino)ethanesulfonic_acid	207.287	462
117	5-aminoindazole	133.152	462
118	(2R)-2,4-dihydroxy-N-(3-hydroxypropyl)-3,3-dimethylbutanamide	205.253	463
119	2-[(carboxymethyl)methylamino]benzoic_acid	209.201	463
120	5-[(2-hydroxy-tert-butyl)amino]-1,3-dihydropyrimidine-2,4-dione	199.209	463
121	6-amino-2-(4-methylpiperazinyl)-3-hydropyrimidin-4-one	209.25	463
122	m-phenylenediamine	108.143	465
123	N-[1-carbamoyl-2-(4-hydroxyphenyl)ethyl]acetamide	222.243	465
124	2-(5,7-dimethyl-2,4,6-trioxo-3,5,7-trihydro-5,7-diazaindol-3-yl)acetic_acid	253.214	466
125	4-methyl-3-(4-pyridyl)-1,2,4-triazole	160.178	466
126	2-hydroxy-2-phenylacetamide	151.165	467
127	N-[7-(methoxymethyl)-5-oxo-4-hydro-1H-1,2,4-triazolino[1,5-a]pyrimidin-2-yl]acetamide	237.218	467
128	N-methylpropylamine	73.1376	467
129	2-(trimethylamino)ethyl_acetate_chloride	146.209	468
130	6-methylpurine	134.14	468
131	3-(2-pyridyl)pentanedioic_acid	209.201	470
132	3-methylpyridazine	94.1158	470
133	3-methyl-3-[(2-oxopropyl)amino]butan-2-one	157.212	471
134	6-amino-5-methyl-3-hydrobenzimidazol-2-one	163.179	471
135	1,3,5-triacetyl-1,3,5-triazaperhydroine	213.236	472
136	1,3-dimethyl-1,3,5-triazaperhydroine-2,4,6-trione	157.129	472
137	N,N-dimethylisopropylamine	87.1644	472
138	N-ethyl-N'-ethylpropane-1,3-diamide	158.2	472
139	3,4-diaminobenzoic_acid	152.152	473
140	4-aminopyrazolo[3,4-d]pyrimidine	135.128	473
141	pyrazine	80.089	473
142	(R)-(-)-2-methylpyrrolidine	85.1486	474
143	2-amino-6-methylhydropyrimidin-4-one	125.13	474
144	2-methylthio-2-imidazoline	116.181	474
145	1,1-dimethylpiperidin-1-ium_chloride	114.21	476
146	ethylenediamine	60.0986	477

147	(4-aminophenyl)-N-methylcarboxamide	150.18	478
148	2-picoline	93.128	478
149	4-azabenzimidazole	119.126	478
150	3,5-di(acetylamino)-2-methylbenzoic_acid	250.254	481
151	methyl_2-{2-[2-(acetylamino)propanoylamino]propanoylamino}propanoate	287.315	481
152	(2E)-3-(dimethylamino)prop-2-enal	99.132	482
153	2-ethylimidazole	96.1316	482
154	2-propylpyridine-4-carboxylic_acid	165.191	482
155	2,6-diaminopyridine	109.13	483
156	3-(diethylamino)propan-1-ol	131.217	483
157	2-aminopyridine	94.1158	484
158	2-(3,5-dimethyl-2,4,6-trioxo(1,3,5-triazaperhydroinyl))-N-methylacetamide	228.207	485
159	2,3-diaminopyridine	109.13	485
160	N-[5-(acetylamino)-2-hydroxyphenyl]acetamide	208.216	485
161	1,3-diacetyl-4,5-dihydroxy-2-oxoimidazolidine	202.166	486
162	1,4-dimethylpiperazine	114.19	486
163	3-piperazinylpropanamide	157.215	487
164	triethyl(methyl)azanium_chloride	116.226	487
165	6-methyl-2-oxohydropyridine-3-carboxylic_acid	153.137	488
166	5-(dimethylamino)-1,3-dihydropyrimidine-2,4-dione	155.156	489
167	quinoxaline-2,3-diol	162.148	490
168	4-aminopiperidine	100.163	492
169	N-(2-pyridyl)acetamide	136.153	492
170	N-[(4,6-dimethylpyrimidin-2-yl)amino]acetamide	180.209	492
171	4-picoline	93.128	493
172	(3-aminophenyl)methan-1-ol	123.154	495
173	N,N-dimethyl-2-(1,3,5-trioxo(4,6,7,3a,7a-pentahydroisoindol-2-yl))acetamide	252.269	495
174	4-(1-amino-2-carboxyethyl)benzoic_acid	209.201	496
175	4-(3-pyridyl)pyrazole	145.163	496
176	5-(aminomethyl)-3-hydrobenzimidazol-2-one	163.179	496
177	amino(iminomorpholin-4-ylmethyl)carboxamidine_chloride	171.202	496
178	isobutylamine	73.1376	496
179	methylpyrazin-2-ylamine	109.13	496
180	7-aminohydroquinoxalin-2-one	161.163	497
181	triethylamine	101.191	497
182	1-(ethylamino)butan-2-ol	117.191	498
183	1-[(tert-butyl)amino]acetone	129.202	499
184	4-amino-4-methylpentan-2-one_oxalic_acid	115.175	499
185	6-aminohexan-1-ol	117.191	499
186	methyl_2-oxopiperidine-3-carboxylate	157.169	499
187	1,2-diaminopropane	74.1254	500
188	methyl_2-(6-oxohydropyridazinyl)acetate	168.152	500
189	piperazine	86.1364	500
190	{2-[3-(hydroxymethyl)-2-pyridyl]-3-pyridyl}methan-1-ol	216.239	501
191	2-(3-pyridyl)imidazole	145.163	502
192	2-methyl-1,2,3,4-tetrahydroisoquinoline-3-carboxylic_acid	191.229	502
193	3-(dimethylamino)-2,2-dimethylpropan-1-ol	131.217	504
194	4-phenylpyridine-2,5-dicarboxylic_acid	243.218	504
195	3-hydroxy-1,2-dimethylhydropyridin-4-one	139.154	505
196	8-methyl-8-azatricyclo[6.2.2.0<2,7>]dodecan-4-ol	196.312	505
197	ethenodeoxyadenosine	275.266	505

198	1-(2-aminoethyl)piperazine	129.205	506
199	1-propylpiperidine-3-carboxamide	170.254	506
200	2,6-lutidine	107.155	506
201	(3S,4R)-4-ethoxyoxolan-3-amine_hydrochloride	131.174	507
202	1-methylpiperazine	100.163	507
203	2-methylpiperazine	100.163	507
204	4-(2-methoxyphenyl)-1,2,4-triazolidine-3,5-dione	207.188	507
205	4-(aminomethyl)piperidine	114.19	507
206	4-(aminomethyl)pyridine	108.143	507
207	N,N,N',N'-tetramethylguanidine	115.178	507
208	1-(2-(2-pyridyl)ethyl)pyrrolidin-2-one	190.244	508
209	N,N-dimethylethylenediamine	88.1522	508
210	1,2,3,4-tetrahydroisoquinoline-3-carboxylic_acid	177.202	509
211	1-ethylpiperidine	113.202	509
212	2,6-diaminotoluene	122.169	509
213	cyclopentylamine	85.1486	509
214	N,N'-dimethylethylenediamine	88.1522	509
215	N-methylethylenediamine	74.1254	509
216	5-amino-1,3-dimethyl-3-hydrobenzimidazol-2-one	177.205	510
217	N-(2-(2-piperidyl)ethyl)acetamide	170.254	510
218	2-(methylamino)pyridine	108.143	511
219	4-(methylamino)pyridine	108.143	511
220	N-(2-aminophenyl)acetamide	150.18	511
221	1-methylhomopiperazine	114.19	512
222	4-[(2-aminophenyl)carbonyl]piperazin-2-one	219.243	512
223	(2S)-1-benzylpyrrolidine-2-carboxylic_acid	205.256	513
224	(6-oxohydropyridazin-3-yl)-N-(3-pyridyl)carboxamide	216.199	513
225	1-dimethylamino-2-propylamine	102.179	513
226	imidazo[5,4-g]benzimidazole	158.162	513
227	(4-aminophenyl)-N,N-dimethylcarboxamide	164.207	514
228	4-aminobenzenecarboxamidine	135.168	514
229	(R)-(+)-3-(dimethylamino)pyrrolidine	114.19	515
230	ethyl_(2R)-5-oxopyrrolidine-2-carboxylate	157.169	515
231	(2-aminophenyl)methan-1-ol	123.154	516
232	2-[(iminomorpholin-4-ylmethyl)amino]-6-methyl-3-hydropyrimidin-4-one	237.261	516
233	2-amino-2-phenylacetamide	150.18	516
234	2-ethyl-4-methylimidazole	110.158	516
235	N'-(3-hydroxypropyl)-N-(methylethyl)ethane-1,2-diamide	188.226	516
236	trans-2,5-dimethylpiperazine	114.19	516
237	2,5-lutidine	107.155	517
238	2-isopropylimidazole	110.158	517
239	2-phenylpiperidine-2-carboxylic_acid	205.256	517
240	3-{N-[4-(acetylamino)phenyl]carbamoyl}propanoic_acid	250.254	517
241	4-(2H-benzo[d]1,3-dioxolen-5-yl)-1,2,4-triazolidine-3,5-dione	221.172	517
242	1-(3-aminopropyl)imidazole	125.173	518
243	1-ethylpiperazine	114.19	518
244	2,3-lutidine	107.155	518
245	ethyl_2-acetyl-3-amino-5-oxo-3-pyrazoline-4-carboxylate	213.193	518
246	N-(5-quinolyl)acetamide	186.213	518
247	o-phenylenediamine	108.143	518
248	2-ethylpyridine	107.155	519
249	5-[((1E)-2-(3-pyridyl)-1-azavinyl)amino]-2H-1,2,4-triazin-3-one	216.202	519

250	2,6-dimethylpiperazine	114.19	520
251	4-(4-pyridyl)butanoic_acid	165.191	520
252	leucinamide_hydrochloride	130.189	520
253	3-(3-aminophenyl)-1-methyl-1,3-diazolidine-2,4-dione	205.216	521
254	N-ethyl-4-pyridinemethylamine	136.196	521
255	2-methylquinoline-4-carboxamide	186.213	522
256	butylamine	73.1376	522
257	1-acetyl-4-(propylamino)piperidine	184.281	524
258	2,2,5-trimethyl-1,2,6-trihydropyridin-3-one	139.197	524
259	3-(3-aminophenyl)-5-methyl-1,3-diazolidine-2,4-dione	205.216	525
260	((1R,2R)-6-azabicyclo[4.4.0]dec-2-yl)methan-1-ol	169.266	526
261	hexamethylenediamine	116.206	526
262	2,4-lutidine	107.155	527
263	3-(2-pyridyl)propylamine	136.196	527
264	4-(2-morpholin-4-ylethyl)morpholine	200.28	527
265	4-(3-aminophenyl)-1,2-dimethyl-1,2,4-triazolidine-3,5-dione	220.23	527
266	aniline	93.128	527
267	decahydroquinolin-4-one	153.224	527
268	1-(3-aminophenyl)-2-methyl-2-imidazolin-5-one	189.216	528
269	1,2,6-trimethylhydropyridin-4-one	137.181	528
270	2-(dimethylamino)pyridine	122.169	528
271	3,4-lutidine	107.155	528
272	pyridino[3,2-h]quinoline-4,7-diol	212.207	528
273	2-(2-pyridyl)ethylamine	122.169	529
274	2-propylpyridine-4-carboxamide	164.207	529
275	quinazoline-4-ylamine	145.163	529
276	2,2-dimethyl-3-(pyrrolidin-1-yl)propanal	155.239	530
277	2,4-diaminotoluene	122.169	530
278	imidazo[5,4-g]quinoxaline	170.173	530
279	2-(aminomethyl)piperidine	114.19	531
280	3-(2-imidazolin-2-yl)phenylamine	161.206	531
281	N'-[3-(dimethylamino)propyl]-N-cyclopropylethane-1,2-diamide	213.279	531
282	3,5-lutidine	107.155	532
283	4,5-dimethoxycyclohexa-3,5-diene-1,2-dione	168.149	532
284	N,N'-diethylethylenediamine	116.206	532
285	N,N-dimethyl-p-phenylenediamine	136.196	532
286	pyrazole	68.078	532
287	2-imidazolin-2-ylpyrazin-2-ylamine	163.182	533
288	4-methylpyrimidine	94.1158	533
289	1-cyclopropyl-2-pyridylethan-1-one_iodide	162.211	534
290	3-(diethylamino)propylamine	130.233	534
291	6-methyl-2-(methylethyl)pyrimidin-4-ol	152.196	534
292	N,N,N',N'',N''-pentamethyldiethylenetriamine	173.301	534
293	N-[(4-hydroxyphenyl)ethyl]acetamide	179.218	534
294	1-allylpiperazine	126.201	535
295	2-(2-methylaminoethyl)pyridine	136.196	535
296	3-ethylpyridine	107.155	535
297	1,3-dimethyl-1,3-diazinan-2-one	128.174	536
298	1-isopropylpiperazine	128.217	536
299	4-ethylpyridine	107.155	536
300	5,6,7-trihydrocyclopenta[2,1-d]pyrimidine-2-ylamine	135.168	536
301	methyl_2-(2,4-diaminophenyl)acetate	180.206	536
302	N-[4,6-di(acetylamino)pyrimidin-2-yl]acetamide	251.244	536

303	1-(N-propylcarbamoyl)piperidine-3-carboxamide	213.279	537
304	4-amino-2,6-dimethylphenol	137.181	537
305	N-(4-aminophenyl)[(4-aminophenyl)amino]carboxamide	242.28	537
306	1,2-dimethylpropylamine	87.1644	538
307	1-acetyl-4-(4-piperidyl)piperidine	210.319	538
308	2-(aminomethyl)benzoic_acid	151.165	538
309	5-amino-3-hydroisobenzofuran-1-one	149.149	538
310	8-methyl-2-morpholin-4-ylimidazolidino[1,2-e]1,3,5-triazin-4-one	237.261	538
311	diallylmethylamine	111.186	538
312	N,N,N-trimethylanilinium_chloride	136.216	538
313	pyrrole-2-carbaldehyde	95.1006	538
314	(3,4-dimethoxyphenyl)-N-(???methyl)carboxamide	253.257	539
315	1-(2-aminoethyl)piperidine	128.217	539
316	2-(pyrrolidinylmethyl)pyridine	162.234	539
317	3,5-dimethylpyrazole	96.1316	539
318	4-(2-(4-pyridyl)ethyl)pyridine	184.24	539
319	N-propylethylenediamine	102.179	539
320	(2-(2-pyridyl)ethyl)pyrimidin-4-ylamine	200.243	540
321	(3-amino(4-pyridyl))dimethylamine	137.184	540
322	1,2-diaminocyclohexane	114.19	540
323	5-aminoindole	132.165	540
324	methyl_3-piperidylpropanoate	171.239	540
325	N,N,N'-triethylethylenediamine	144.259	540
326	pyrazin-2-yl-N-(3-pyridyl)carboxamide	200.199	540
327	(2S)-2-amino-2-phenylethan-1-ol	137.181	541
328	1-(4-aminophenyl)pyrrolidin-2-one	176.218	541
329	2-(4-oxo-3-hydroquinazolin-3-yl)acetic_acid	204.185	541
330	2,7-dimethylimidazo[5,4-g]benzimidazole	186.216	541
331	m-xylylenediamine	136.196	541
332	N,N'-dimethyl-1,6-hexanediamine	144.259	541
333	N,N-dimethylbutylamine	101.191	541
334	o-xylylenediamine	136.196	541
335	1-ethylpropylamine	87.1644	542
336	6-quinolylmethylamine	158.202	542
337	1-(aminomethyl)cyclohexan-1-ol	129.202	543
338	2-amino-5-diethylaminopentane	158.286	543
339	4-morpholin-4-ylbenzene-1,3-diamine	193.248	543
340	N-ethylbutylamine	101.191	543
341	2-amino-3-phenylbutanoic_acid_chloride	179.218	544
342	3-(2-hydroxyethyl)-1,3-dihydroquinazoline-2,4-dione	206.201	544
343	3,6-dimethyl-1-(3-pyrrolidinylpropyl)-1,3-dihydropyrimidine-2,4-dione	251.328	544
344	4-(3-(4-pyridyl)propyl)pyridine	198.267	544
345	5-methyl-4-azabicyclo[4.4.0]decan-1-ol	169.266	544
346	benzamide	121.138	544
347	methyl_3-(3-pyridyl)propanoate	165.191	544
348	N-(2-pyridylmethyl)[(4-{[(2-pyridylmethyl)amino]carbonylamino}butyl)amino]carboxamide	356.427	544
349	N-methyl-2-(3-oxo(1,2,4-trihydroquinoxalin-2-yl))acetamide	219.243	544
350	2-(phenylcarbonylamino)acetic_acid	179.175	545
351	2,3-dimethylquinoxaline-6-ylamine	173.217	545
352	2-aminopentane	87.1644	545
353	5-methyl-5-azabicyclo[4.4.0]decan-2-one	167.25	545
354	benzimidazole	118.138	545

355	bis(3-pyridylmethyl)amine	199.255	545
356	methyl[(1-methylpyrrol-2-yl)methyl]amine	124.185	545
357	quinoxaline-6-ylamine	145.163	545
358	(4S,5R)-2,4,5-tri(4-pyridyl)-2-imidazoline	301.35	546
359	1-methyl-3-[4-(2-pyrrolidinylacetyl)piperazinyl]azolidine-2,5-dione	308.38	546
360	4-(pyrrolidinylmethyl)phenylamine	176.261	546
361	4,4-dimethylazaperhydroine-2,6-dione	141.169	546
362	N-((1S,6R)-5-methyl-5-azabicyclo[4.4.0]dec-2-yl)acetamide	210.319	546
363	(3S)-3-amino-2,4-dimethylpentan-2-ol	131.217	547
364	4-methoxy-6-pyrrolidinyl-1H-1,3,5-triazin-2-one	196.208	548
365	6-[4-(4,6-diamino-1,3,5-triazin-2-yl)phenyl]-1,3,5-triazine-2,4-diamine	296.294	548
366	N-{4-[N-(2-pyridylmethyl)carbamoyl]phenyl}acetamide	269.302	548
367	N-isopropyl-N-methyl-tert-butylamine	129.245	548
368	trans-N,N'-dimethylcyclohexane-1,2-diamine	142.244	548
369	amino(imino-1,4-thiazaperhydroin-4-ylmethyl)carboxamidine	187.262	549
370	2-ethyl-4-methylspiro[2,4,5,6,3a,6a-hexahydro-2,5-diazapentalene-6,3'-indoline]-1,3,10-trione	299.329	550
371	N,N-diisopropylethylenediamine	144.259	550
372	5-[3-(dimethylamino)propyl]-5-azabicyclo[4.4.0]decan-2-ol	240.388	551
373	1-(2-piperazinylacetyl)-4-piperidylpiperidine-4-carboxamide	337.464	552
374	4-amino-6-(tert-butyl)-1-methyl-1,3,5-triazin-2-one	182.225	552
375	cis-1,8-diamino-p-menthane	170.297	552
376	N,N,N',N'-tetraethylethylenediamine	172.313	552
377	N,N,N',N'-tetramethyl-1,3-propanediamine	130.233	552
378	[4-(2-aminoethyl)phenyl]diethylamine	192.303	553
379	1,8-diaminooctane	144.259	553
380	1-acetyl-5-amino-2-methylindoline	190.244	553
381	2-(cyclohexylamino)ethan-1-ol	143.228	553
382	3-morpholin-4-ylphenylamine	178.233	553
383	methyl_6-aminohexanoate	145.201	553
384	N,N,N',N'-tetraethyl-1,3-propanediamine	186.34	553
385	[2-(2-amidinothioethylthio)ethyl]thiocarboxamidine	238.383	554
386	3-[5-(carboxymethyl)-3-oxo-5-hydro-1,2,4-triazino[2,3-a]benzimidazol-2-yl]propanoic_acid	316.273	554
387	(2S,1R)-1-aminoindan-2-ol	149.192	555
388	1-[(3,4-dihydroxyphenyl)methyl]-1,2,3,4-tetrahydroisoquinoline-6,7-diol_bromide	287.315	555
389	3-amino-1,3,4-trihydroquinolin-2-one	162.191	555
390	ethyl_5-(amidinoamino)-2-aminopentanoate_chloride	202.256	555
391	phenylurea	136.153	555
392	(1-methyl(4-piperidyl))(3-pyridylmethyl)amine	205.302	556
393	1,5-dimethyl-3-(2-oxopropyl)-1,3,5-triazaperhydroine-2,4,6-trione	213.193	556
394	1-{[(2-phenylethyl)(hydroxyphosphoryl)]methyl}pyrrolidin-2-one	267.264	556
395	1-methylbenzimidazole	132.165	556
396	5-(3,5-dimethylpyrazolyl)-1H-1,2,3,4-tetraazole	164.169	556
397	aminodimethylbenzylamine	151.231	556
398	indoline	119.166	556
399	(2-methylpropyl)(2-morpholin-4-ylethyl)amine	186.297	557
400	1-(4-pyridylmethyl)-4-(2-pyridylmethyl)piperazine	268.361	557
401	7-azaindole	118.138	557
402	ethyl_2-(1,4-dimethylpiperazin-2-yl)acetate	200.28	557
403	N,N'-diisopropyl-1,3-propanediamine	158.286	557
404	N-methylaniline	107.155	557
405	pyridino[3,2-h]quinoline-5,6-dione	210.192	557

406	quinoline	129.161	557
407	2,3-diaminobenzoic_acid	152.152	558
408	3-(carboxymethyl)indole-2,5-dicarboxylic_acid	263.206	558
409	3-amino-1-(3-methylpiperidyl)propan-2-ol	172.27	558
410	4-quinolylamine	144.176	558
411	N-[(2,2,6,6-tetramethyl(4-piperidylidene))azamethyl]-N'-[(2,2,6,6-tetramethyl(4-piperidylidene))azamethyl]ethane-1,2-diamide	392.543	558
412	1-[4-amino-3-(2-oxopyrrolidinyl)phenyl]pyrrolidin-2-one	259.307	559
413	2,2-diprop-2-enylpiperazine_chloride_chloride	166.266	559
414	2-(2-pyrrolidinylacetylamino)benzamide	247.296	560
415	2,3-diaminotoluene	122.169	560
416	2-amino-1-phenylethan-1-one	135.165	560
417	((3S)-3-1,2,3,4-tetrahydroisoquinolyl)methan-1-ol	163.219	562
418	[1-(3-methoxypropyl)pyrrolidin-3-yl]methanamine	172.27	562
419	1H-2,4,5,6,7-pentahydroindazol-3-one	138.169	562
420	2-methylbenzimidazole	132.165	562
421	amino[(4-oxocyclohexa-2,5-dienylidene)azamethyl]carboxamidine	164.166	562
422	N,N-diethyl(4-methylpiperazinyl)carboxamide	199.295	562
423	N,N-dimethyldipropylenetriamine	159.274	562
424	5-amino-2-methylbenzoic_acid	151.165	563
425	heptamethyleneimine	113.202	563
426	N-(oxolan-2-ylmethyl)[(4-{[(oxolan-2-ylmethyl)amino]carbonylamino}butyl)amino]carboxamide	342.437	563
427	5-amino-1,3,3-trimethylcyclohexanemethylamine	170.297	564
428	N-[(2-hydroxyindol-3-ylidene)azamethyl]acetamide	203.2	564
429	2-(2-oxoindolin-3-yl)acetic_acid	191.186	565
430	3-(4-methylquinolyl)propanoic_acid_bromide	216.259	565
431	bis(hexamethylene)triamine	215.381	565
432	N-(4-hydroxyphenyl)-N-methylacetamide	165.191	565
433	2,4-dimethylquinoline-7-carboxylic_acid	201.224	566
434	2,5-dimethylpyrrole-3,4-dicarbaldehyde	151.165	566
435	3,4-diaminotoluene	122.169	566
436	N-(2-aminoethyl)-1,3-propanediamine	117.194	566
437	1-(4-hydroxyphenyl)pyrrolidin-2-one	177.202	567
438	2-(acetylamino)-3-methylpentanoic_acid	173.211	567
439	3-(2-aminoethyl)-3-hydroquinazolin-4-one	189.216	567
440	N-(4-aminophenyl)butanamide	178.233	567
441	N-[4-(4-pyridylmethyl)phenyl]acetamide	226.277	567
442	N-benzylmethylamine	121.182	567
443	2,6-dimethylpyrazine	108.143	568
444	2-amino-N-phenylacetamide	150.18	568
445	3,3'-diamino-N-methyldipropylamine	145.247	568
446	5,6-dimethoxyindolin-2-one	193.202	568
447	N-(2-hydroxyethyl)-2-(4-methyl-1-oxo(2-hydrophthalazin-2-yl))acetamide	261.28	568
448	N-allylcyclopentylamine	125.213	568
449	N-methylcyclohexylamine	113.202	568
450	(4-pyridylmethyl)(2,2,6,6-tetramethyl(4-piperidyl))amine	247.383	569
451	2-(pyrrolidinylmethyl)benzoic_acid	205.256	569
452	2-[(N-phenylcarbamoyl)amino]acetic_acid	194.19	569
453	3-[(1E)-3,3-bis(hydroxymethyl)-4-hydroxy-2-azabut-1-enyl]-4-hydroxy-6-methylhydroquinolin-2-one	306.318	569
454	methyl_3-(3-piperidyl)propanoate	171.239	569

455	N-(2,2,6,6-tetramethyl(4-piperidyl))-N'-(2,2,6,6-tetramethyl(4-piperidyl))ethane-1,2-diamide	366.546	569
456	N,N'-di-tert-butyl-ethylenediamine	172.313	569
457	1-butyl-1-methylpyrrolidin-1-ium_chloride	156.29	570
458	2,5-dimethylpyrazine	108.143	570
459	2-methyl-4-[(4-methylpiperazinyl)carbonyl]-2-hydrophthalazin-1-one	286.333	570
460	4-[2-(methylpiperidyl)ethylidene]-1,3-dioxolane_iodide	198.285	570
461	methyl_2-(3-amino-4,6-dimethyl-2-oxohydropyridyl)acetate	210.232	570
462	N-(3-pyridylmethyl)[(3-{[N-(3-pyridylmethyl)carbamoyl]amino}phenyl)amino]carboxamide	376.417	570
463	N,N-dimethyl-N'-ethylethylenediamine	116.206	570
464	1,10-diaminodecane	172.313	571
465	3,3'-iminobis(N,N-dimethylpropylamine)	187.328	571
466	5-[3-(dimethylamino)propyl]-5-azabicyclo[4.4.0]decan-2-one	238.372	571
467	1-butylimidazole	124.185	572
468	4,6-dimethylpyrimidine	108.143	572
469	2-imino-5-hydro-3H-1,3,5-triazino[6,1-b]benzoxazol-4-one	202.172	573
470	3-pyrazin-2-yl-5-(2-pyridyl)-1H-1,2,4-triazole	224.224	573
471	amino-N-methyl-N-[2-(phenylamino)ethyl]amide	193.248	573
472	N-((1S)-1-carbamoyl-2-indol-3-ylethyl)acetamide	245.28	573
473	N,N-dimethylbenzylamine	135.208	573
474	triethylenetetramine	146.235	573
475	1,7-dimethylquinoline	158.223	574
476	1-butylpyrrolidine	127.229	574
477	3-(1,3-dioxan-2-yl)phenylamine	179.218	574
478	3-(carboxymethyl)indole-2,6-dicarboxylic_acid	263.206	574
479	4-(aminomethyl)-1-methyl-1,3,4-trihydroquinolin-2-one	190.244	574
480	N-(5,7-dimethylpyridino[3,2-e]pyridin-2-yl)acetamide	215.254	574
481	piperidine-1-carbaldehyde	113.159	574
482	1-(piperidin-4-yl)butan-1-one	155.239	575
483	1,2,3,4-tetrahydroquinoline	133.193	575
484	1-benzylpiperidine-4-carbaldehyde	203.283	575
485	2-ethylbenzimidazole	146.191	575
486	6-aminoindole	132.165	575
487	N-((1E)-2-(3-pyridyl)-1-azavinyl)-2-pyridylcarboxamide	226.237	575
488	N,N-dimethyl-o-toluidine	135.208	575
489	N-acridin-9-ylacetamide	236.273	575
490	4-[benzylamino]butan-1-ol	179.261	576
491	N,N-dimethylaniline	121.182	576
492	N,N-dimethylcyclohexylamine	127.229	576
493	(4-aminophenyl)-N-(cyclopentylideneazamethyl)carboxamide	217.27	577
494	[(2-hydroxyethyl)methylamino]-N-benzamide	194.233	577
495	1-ethyl-4-pyrazol-3-ylpyrazole	162.194	577
496	2-(ethylamino)-4-methylphenol	151.208	577
497	3-[4-(acetylamino)phenyl]propanoic_acid	207.229	577
498	methyl_2-amino-2-phenylacetate	165.191	577
499	N-(2,2,6,6-tetramethyl(4-piperidyl)){2-[N-(2,2,6,6-tetramethyl(4-piperidyl))carbamoyl]phenyl}carboxamide	442.643	577
500	N-(2-amino-4,6-dimethylphenyl)acetamide	178.233	577
501	o-toluidine	107.155	577
502	[2-(pyrrolidinylmethyl)phenyl]methan-1-ol	191.272	578
503	1,1,4,7,10,10-hexamethyltriethylenetetramine	230.396	578
504	1-[3-(methylamino)propyl]-3-hydrobenzimidazol-2-one	205.259	578
505	N,N',N''-trimethylbis(hexamethylene)triamine	257.462	578

506	(indol-3-ylmethyl)dimethylamine	174.245	579
507	1-(2-aminoethyl)-3-methyl-1,3-dihydroquinazoline-2,4-dione	219.243	579
508	1,1,2,3-tetramethylpiperidin-4-one	156.247	579
509	4-pyrrolidinylpyridine	148.207	579
510	methyl_2-(4-oxo-2-piperidyl-3,5,6-trihydropyrimidin-5-yl)acetate	253.3	579
511	p-toluidine	107.155	579
512	[2-(5,5-dimethyl(1,3-dioxolan-4-ylidene))ethyl]trimethylamine_iodide	186.273	580
513	2-prop-2-enylisoquinoline_bromide	170.233	581
514	3-acetyl-4-methylpyrrole	123.154	581
515	4,5-dimethoxy-2-morpholin-4-ylphenylamine	238.286	581
516	4-[4-(dimethylamino)phenyl]butan-2-ol	193.288	581
517	1,4-bis[(3-methyl(2-pyridyl))methyl]piperazine	296.414	582
518	1-[(3,4-dimethoxyphenyl)methyl]-4-piperidylamine	250.34	582
519	2-(phenylcarbonylamino)propanoic_acid	193.202	582
520	2-methylindoline	133.193	582
521	2-[(2-carboxyethyl)amino]benzene-1,4-dicarboxylic_acid	253.211	583
522	N-ethylaniline	121.182	583
523	N-ethylbenzylamine	135.208	583
524	2-aminobenzimidazole	133.152	584
525	N-(4-aminophenyl){3-[N-(4-aminophenyl)carbamoyl]phenyl}carboxamide	346.388	584
526	methyl(2-(2-pyridyl)ethyl)(2-pyridylmethyl)amine	227.308	585
527	N-(2,2,6,6-tetramethyl(4-piperidyl))-2-[(2,2,6,6-tetramethyl(4-piperidyl))amino]acetamide	352.562	585
528	[4-({[3-(dimethylamino)propyl]amino}methyl)phenyl]diethylamine	263.425	586
529	1-(phenylcarbonyl)-4-piperidylpiperidine-4-carboxamide	315.414	586
530	1,2,3,4-tetrahydroisoquinoline	133.193	586
531	1,2-bis(3-aminopropylamino)ethane	174.289	586
532	2-(cyclohexylcarbonylamino)acetic_acid	185.222	586
533	2,3,5,6-tetramethylpyrazine	136.196	586
534	4-(2,4,6-trimethylphenyl)-1,3-thiazolin-2-imine_bromide	218.316	586
535	N,N-diethylbutylamine	129.245	586
536	2-ethylpyrazine	108.143	587
537	5-methylbenzimidazole	132.165	587
538	(4,6-dimethylpyrimidin-2-yl)(5-methyl(4H-1,2,4-triazol-3-yl))amine	204.234	588
539	3-aminoindolin-2-one	148.164	588
540	4-(3-methyl-5-oxo-2-pyrazolinyl)benzoic_acid	218.212	588
541	8-methoxy-2-methylquinoline	173.214	588
542	N-(2-aminophenyl)[(2-aminophenyl)amino]carboxamide	242.28	588
543	1,3-dimethyl-5-[(propylamino)methyl]-3-hydrobenzimidazol-2-one	233.313	589
544	1,3-dimethylbutylamine	101.191	589
545	methyl_3-(N-{3-[N-benzylcarbamoyl]propanoylamino}carbamoyl)propanoate	335.359	589
546	2-pyridyl-N-(3-pyridyl)carboxamide	199.212	590
547	2-(1,5-dimethylpyrazol-4-yl)benzimidazole	212.254	591
548	3-(3-pyridyl)-1,3-dihydroquinazoline-2,4-dione	239.233	591
549	N-[1,1-bis(hydroxymethyl)-2-hydroxyethyl]-N'-phenylethane-1,2-diamide	268.269	591
550	1-aminoindan	133.193	592
551	2-{[2-(3,4-dimethoxyphenyl)ethyl]amino}-3-hydropyrimidin-4-one	275.307	592
552	2-methylbenzylamine	121.182	592
553	5,6-dimethylquinoxaline-2,3-diol	190.201	592
554	cycloheptylamine	113.202	593
555	isoquinolylmethylamine	158.202	593

556	tris[2-(isopropylamino)ethyl]amine	272.476	593
557	[(1E)-1-amino-2-(4,6-dimethylpyrimidin-2-yl)-2-azavinyl](3-pyridylmethyl)amine	256.31	594
558	1-ethyl-3-methyl-4-pyrazol-3-ylpyrazole	176.221	594
559	2-(4-pyridyl)quinoline-4-carboxylic_acid	250.256	594
560	2-hydroxy-2-(6-hydroxy-2-oxocyclohex-1(6)-enyl)-2-hydrocyclopenta[1,2-a]benzene-1,3-dione	272.257	594
561	cyclohexylurea	142.2	594
562	[(2,3-dimethoxyphenyl)methyl]methylamine	181.234	595
563	1-((4aS,9bR)-2,8-dimethylpiperidino[4,3-b]indolin-5-yl)-2-piperidylethan-1-one	327.469	595
564	2-piperidylphenol	177.246	595
565	4-{2-oxo-2-[4-benzylpiperazinyl]ethyl}-2,5-diazabicyclo[4.4.0]decan-3-one	370.494	595
566	2-(3-pyridyl)benzimidazole	195.223	596
567	3-methylbenzylamine	121.182	596
568	4-methylbenzylamine	121.182	596
569	cyclohexanecarboxamide	127.186	596
570	[(4R,5R)-5-(N,N-dimethylcarbamoyl)-2,2-dimethyl(1,3-dioxolan-4-yl)]-N,N-dimethylcarboxamide	244.29	597
571	(S)-(-)-N,alpha-dimethylbenzylamine	135.208	598
572	3-(3-hydroxypropyl)indolin-2-one	191.229	598
573	6-benzyl-1,3,5-triazine-2,4-diamine	201.23	598
574	8-methyl-3-phenyl-1,3,4,8-tetraazaspiro[4.5]decan-2-one	246.311	598
575	N-(1-methyl-2-oxo-2-piperidylethyl)-3-pyridylcarboxamide	261.323	598
576	N-(3,5-diamino-2-methylphenyl)benzamide	241.292	598
577	propyl_piperidine-2-carboxylate	171.239	598
578	allylcyclohexylamine	139.24	599
579	3-(pyrrolylmethyl)pyridine	158.202	600
580	N-((1E)-2-(4-pyridyl)-1-azavinyl)benzamide	225.249	600
581	N-(3-(2-imidazolin-2-yl)phenyl)[(3-(2-imidazolin-2-yl)phenyl)amino]carboxamide	348.407	600
582	2,5-dimethylbenzimidazole	146.191	601
583	methyl_2-(2,2,6,6-tetramethyl-4-piperidyl)acetate	213.319	601
584	N-methyl-phenethylamine	135.208	601
585	phenylpiperazine_chloride	162.234	601
586	1-cyclohexyl-3-(2,2,6,6-tetramethyl(4-piperidyl))imidazolidin-4-one	307.478	602
587	4,5-dimethyl-1,2-phenylenediamine	136.196	602
588	[4-amino-6-(dimethylamino)(1,3,5-triazin-2-yl)]dimethylamine	182.228	603
589	4-phenyl-1-(4-piperidyl)-4-imidazolin-2-one	243.308	603
590	N-(3-acetylphenyl)acetamide	177.202	603
591	2-(4-methyl-1-oxo(2-hydrophthalazin-2-yl))-N-(2-pyridylmethyl)acetamide	308.339	604
592	2,2,4-trimethyl-1,2,3,4-tetrahydroquinolin-6-ol	191.272	604
593	3-(4-pyridylmethyl)benzo[d]1,2,3-triazin-4-one	238.248	604
594	5,5-dimethyl-2-{[(2-piperazinylethyl)amino]ethylidene}cyclohexane-1,3-dione	293.408	604
595	quinoxaline-2-carboxamide	173.174	604
596	2-(2-piperazinylethyl)benzo[c]azolidine-1,3-dione	259.307	605
597	3-methyl-N-methylbenzylamine	135.208	605
598	methyl(2-phenoxyethyl)amine	151.208	605
599	3,5-bis(2-oxopyrrolidinyl)benzoic_acid	288.302	606
600	diisobutylamine	129.245	606
601	N-isopropylaniline	135.208	606
602	1,1-diethoxy-2-piperidylethane	201.308	607

603	2,2,4,4-tetramethyl-3-pentanone_imine	141.256	607
604	2,3-dimethylaniline	121.182	607
605	4-methylpyridino[3,2-h]quinoline	194.235	607
606	N-(3,4,5-trimethoxyphenyl)acetamide	225.244	607
607	2-(3-methoxyphenoxy)ethylamine	167.207	608
608	5-benzyl-1H-1,2,3,4-tetraazole	160.178	608
609	8-methoxy-1,2,3,4-tetrahydroquinoline	163.219	608
610	(6-(1H-1,2,3,4-tetraazin-2-yl)-4-amino(1,3,5-triazin-2-yl))diethylamine	249.278	609
611	10-methyl-3,10-dihydrobenzo[g]pteridine-2,4-dione	228.21	609
612	2-aminobenzoic_acid	137.138	609
613	7-aminoindole	132.165	609
614	cyclohexanemethylamine	113.202	609
615	1,2-bis(2-morpholin-4-yl-2-oxoethyl)-4-phenyl-1,2,4-triazolidine-3,5-dione	431.447	610
616	1-methoxy-2-(2-piperazinylethoxy)benzene	236.313	610
617	3-(carboxymethyl)-2-methylindole-5-carboxylic_acid	233.223	610
618	N,N-dimethyl-p-toluidine	135.208	610
619	3-methyl-4-phenyl-1,2,4-triazole	159.19	611
620	7-amino-4,8-dimethylhydroquinolin-2-one	188.229	611
621	(4-methylphenyl)imidazole	158.202	612
622	3,3-dimethyl-2-phenyl-1-pyrrolin-5-ol	189.257	612
623	4-aminoindan	133.193	612
624	5-methyl-1-(3-(1,2,3,4-tetraazolyl)phenyl)-1,2,3,4-tetraazole	228.216	612
625	6-(methylpropyl)-1-benzyl-1,3,5,6,7,8-hexahydro-6,8-diazaquinazoline-2,4-dione	314.386	612
626	dimethyl(4-{[(2,2,6,6-tetramethyl(4-piperidyl))amino]methyl}phenyl)amine	289.463	612
627	propyl_3-[(3,5-dioxo-2H,4H-1,2,4-triazin-6-yl)amino]propanoate	242.234	612
628	tripropylamine	143.272	612
629	2-methoxy-3-methylbenzenecarbohydrazide	180.206	613
630	3-phenoxypropylamine_chloride	151.208	613
631	N-(2-methoxyethyl)-2-(3-methyl-2-oxo(3-hydrobenzimidazolyl))acetamide	263.296	614
632	N-{(1E)-2-[4-(dimethylamino)phenyl]-1-azavinyl}-2-(3,5-dioxo(2H,4H-1,2,4-triazin-6-yl))acetamide	316.319	614
633	1,12-diaminododecane	200.367	615
634	1-acetylindoline-5-carboxylic_acid	205.213	615
635	5-amino-1,3-dimethyl-6-piperidyl-3-hydrobenzimidazol-2-one	260.338	615
636	N-methyl-o-toluidine	121.182	615
637	1,3-dimethyl-2,4-dioxo-1,3-dihydroquinazoline-6-carboxylic_acid	234.211	616
638	1-(3,4-dimethylphenyl)-6-(methylpropyl)-1,3,5,6,7,8-hexahydro-6,8-diazaquinazoline-2,4-dione	328.413	617
639	3-methyl-5-oxo-1-phenyl-3-pyrazoline-4-carbaldehyde	202.212	617
640	4-methoxy-2,3-dimethylphenylamine	151.208	617
641	4-methylpiperidyl_3-piperidyl_ketone	210.319	617
642	2-(butylamino)-N-{4-[2-(butylamino)acetylamino]phenyl}acetamide	334.461	618
643	3,5-dimethyl-9-(2-phenylethyl)-3,5,7,9-tetraazabicyclo[4.4.0]decane-2,4-dione	302.375	618
644	4-vinylaniline	119.166	618
645	5-(cyclohexylamino)-6-methyl-2H-1,2,4-triazin-3-one	208.263	618
646	3-{[2-(dimethylamino)ethyl]amino}-5-phenylcyclohex-2-en-1-one	258.363	619
647	methylethyl_2-[(3,5-dioxo-2H,4H-1,2,4-triazin-6-yl)amino]propanoate	242.234	619
648	2,4,8-trimethylquinoline	171.241	620
649	N,N-dimethyl-m-toluidine	135.208	620
650	triamterene	253.266	620

651	3-(2-aminoethyl)-4,6-dimethyl-3-hydropyrimidin-2-one	167.21	621
652	5-aminoindan	133.193	621
653	N-(2-hydroxypropyl)[1-(4-methylphenyl)-5-oxopyrrolidin-3-yl]carboxamide	276.335	621
654	[3-(2,5-dioxoazolidinyl)phenyl]-N-(2-pyridyl)carboxamide	295.297	622
655	1,2-dimethylindole-5-ylamine	160.218	622
656	2-(4-methyl-1-oxo(2-hydrophthalazin-2-yl))-N-(2,2,6,6-tetramethyl(4-piperidyl))acetamide	356.467	622
657	2,9-dimethylpyridino[3,2-h]quinoline	208.262	622
658	3,5-dimethylaniline	121.182	622
659	acridine-4-ylamine	194.235	622
660	N-(4-methylphenyl)-3-(4-methylpiperazinyl)propanamide	261.366	622
661	1-(2-phenylethyl)piperidine-3-carboxylic_acid	233.31	623
662	[4-(2-aminoethyl)phenyl]dimethylamine	164.25	624
663	2-(trimethylamino)ethyl_benzoate_chloride	208.28	624
664	4-(N-phenylcarbamoyl)butanoic_acid	207.229	624
665	acridine-9-ylamine	194.235	624
666	N-[(1E,3E)-4-(2-pyridylcarbonylamino)-1,4-diazabuta-1,3-dienyl]-2-pyridylcarboxamide	296.288	624
667	1-(cyclohexylcarbonyl)-4-piperidylpiperidine-4-carboxamide	321.462	625
668	3-amino-4-{[2-(diethylamino)ethyl]amino}chromen-2-one	275.35	625
669	3-oxo-N-benzylbutanamide	191.229	625
670	5,5-dimethyl-3-[(1,2,3,4-tetrahydroisoquinolylmethyl)amino]cyclohex-2-en-1-one	284.4	625
671	2-(2-methylpyridyl)-1-phenylethan-1-one_chloride	212.271	626
672	3-azatetracyclo[7.6.1.0<2,7>.0<10,15>]hexadeca-2(7),3,5,10(15),11,13-hexaen-1-ol	223.274	626
673	ethyl_4,6-dimethyl-2-oxo-3,4,5-trihydropyrimidine-5-carboxylate	198.221	626
674	N-(carbamoylmethyl)-2-quinolylcarboxamide	229.238	626
675	N-{(1E)-2-[5-(hydroxymethyl)-2,4-dimethylphenyl]-1-azavinyl}aminoamide	221.258	626
676	N-ethyl-m-toluidine	135.208	626
677	1-(tert-butyl)-2,3-dimethylpiperidin-4-one	183.293	627
678	3-ethyl-9-hydroxy-2-methyl-5-hydropyridino[1,2-a]pyrimidin-4-one	204.228	627
679	4-(dimethylamino)-2-methyl-1-phenylbutan-2-ol	207.315	627
680	benzyltriethylazanium_chloride	192.324	627
681	2-pyrrolylpropanoic_acid	139.154	628
682	3-hydroxy-4-(iminoethyl)-1-methyl-5-phenyl-3-pyrrolin-2-one	230.266	628
683	4-amino-3-(3,4-dimethoxyphenyl)azoline-2,5-dione	248.238	628
684	4-methyl-2-[3-(methylamino)propyl]phenol	179.261	628
685	7,7-dimethyl-3-benzyl-1,2,3,4,6,7,8-heptahydroquinazolin-5-one	270.374	628
686	8-amino-6-methyl-1,3,4-trihydroquinolin-2-one	176.218	628
687	dibutylamine	129.245	628
688	1-phenyl-2-pyridylpropan-1-one	212.271	629
689	2-pyridyl-N-[2-(2-pyridylcarbonylamino)ethyl]carboxamide	270.29	629
690	N-ethyl-o-toluidine	135.208	629
691	(2-methylphenyl)-N-(2,2,6,6-tetramethyl(4-piperidyl))carboxamide	274.405	630
692	[4-(4-aminophenyl)piperazinyl]-N,N-diethylcarboxamide	276.381	630
693	5-quinolylhydrazine	159.19	630
694	6-[(2,3-dimethylphenyl)amino]-1,3-dihydropyrimidine-2,4-dione	231.254	630
695	methyl_6-(methoxycarbonyl)pyridine-2-carboxylate	195.174	630
696	(phenylamino)-N-(2,4,6-trimethyl(3-pyridyl))carboxamide	255.319	631
697	1-benzyl-4-pyridylamine_bromide	185.248	631
698	3,3-dimethyl-1-phenylurea	164.207	631

699	N-cyclohexylformamide	127.186	631
700	hexyltrimethylazanium_bromide	144.279	632
701	2-indol-3-ylbutanedioic_acid	233.223	633
702	3-[(2-methylpropyl)amino]-3-phenylpropanoic_acid	221.299	633
703	N-(4-methoxy-2,6-dimethylphenyl)acetamide	193.245	633
704	2-acetyl-3-methyl-4-oxo-5,6,7-trihydroindole	191.229	634
705	5-phenyl-5-hydro-1,2,3-triazolo[4,5-d]pyridazin-4-ol	213.198	634
706	N'-(2-methoxyethyl)-N-{[3-({[N-(2-methoxyethyl)carbamoyl]carbonylamino}methyl)phenyl]methyl}ethane-1,2-diamide	394.427	634
707	N-ethyl[4-(2-1,2,3,4-tetrahydroisoquinolylmethyl)phenyl]carboxamide	294.396	634
708	(3,7-dimethyl(2-quinolyl))(3-pyridylmethyl)amine	263.341	636
709	2-(1,3-dimethyl-2,5-dioxo(1,3-diazolidin-4-yl))-N-cyclohexylacetamide	267.327	636
710	cyclohexyl_2-[(2,2,6,6-tetramethyl-4-piperidyl)amino]acetate	296.452	636
711	indol-6-ylmethan-1-ol	147.176	636
712	1-[2-(2-methoxyphenoxy)acetyl]imidazolidin-2-one	250.254	637
713	phenylpurin-6-ylamine	211.226	637
714	N,N-dimethylhexylamine	129.245	638
715	2,3-dimethyl-6-phenylpyridine	183.252	639
716	3-methylcyclohex-2-en-1-one	110.155	639
717	4-((1E)-2-(4-pyridyl)-1-azavinyl)-2,3-dimethyl-1-phenyl-3-pyrazolin-5-one	292.34	639
718	2-amino-N-[2-(methylethyl)phenyl]acetamide	192.26	640
719	6-cyclopentyl-1-(2-phenylethyl)-1,3,5,6,7,8-hexahydro-6,8-diazaquinazoline-2,4-dione	340.424	640
720	N'-(2-methylpropyl)-N-(2,2,6,6-tetramethyl(4-piperidyl))ethane-1,2-diamide	283.413	640
721	rolitetracycline	527.573	640
722	4-{[4-(dimethylamino)phenyl]azamethylene}cyclohexa-2,5-dien-1-one	226.277	641
723	benzyl-2-pyridylamine	184.24	642
724	dimethyl[(4-vinylphenyl)methyl]amine	161.246	642
725	4-methyl-2-{4-[(4-methylpiperazinyl)carbonyl]phenyl}-2-hydrophthalazin-1-one	362.43	643
726	N-methyldibutylamine	143.272	643
727	4-(hydroxymethylene)-3-methyl-1-(4-methylphenyl)-1,2-diazolin-5-one	216.239	644
728	5-acetyl-1,3-dimethyl-2-oxo-3-hydrobenzimidazole	204.228	644
729	diethyl_2-{[(acetylamino)amino]aminomethylene}propane-1,3-dioate	259.261	644
730	2H,4H-benzo[e]1,4-oxazin-3-one	149.149	645
731	2-phenylbenzimidazole	194.235	645
732	3,4-diacetyl-1,2,5-trimethylpyrrole	193.245	645
733	methyl_2-(1,3,7-trimethyl-2,6,8-trioxo-1,3,7-trihydropurin-9-yl)acetate	282.255	645
734	N-((1E)-2-phenyl-1-azavinyl)-4-pyridylcarboxamide	225.249	645
735	N-[(1E)-2-(2-methylindol-3-yl)-1-azavinyl]-3-(3-methyl-5-oxo(2-pyrazolin-4-yl))propanamide	325.369	645
736	phenyl(1,2,5-trimethyl(4-piperidyl))amine	218.341	645
737	2-amino-4-propylphenol	151.208	646
738	2-oxoindoline-3-carbaldehyde	161.16	646
739	3-phenylthiopropylamine	167.268	647
740	4-(4-pyridylmethyl)phenylamine	184.24	647
741	dimethyl{3-[(5,6,7,8,9-pentahydro-4aH-carbazol-3-ylmethyl)amino]propyl}amine	285.431	647
742	(1R)-1-[((10S,11aS,9R)-9-ethyl-2,3-dimethoxy(5,6,7,11a-tetrahydropiperidino[2,1-a]isoquinolin-10-yl))methyl]-6,7-dimethoxy-1,2,3,4-tetrahydroisoquinoline_chloride_hydrate	480.646	648
743	2-[2-(2-aminophenoxy)ethoxy]phenylamine	244.293	648

744	2-methylbenzo[d]1,3-oxazin-4-one	161.16	649
745	methyl_3-(3-amino-4-methylphenyl)propanoate	193.245	649
746	1-(2-methylindolinyl)-2-(4-methylpiperazinyl)ethan-1-one	273.377	650
747	3-[(1E,3E)-4-(3-aminophenyl)-1-methyl-2,3-diazapenta-1,3-dienyl]phenylamine	266.345	650
748	amino-N-[2-(2-methylindol-3-yl)ethyl]amide	217.27	650
749	quinoline-5-carbaldehyde	157.171	650
750	2-((1E)prop-1-enyl)-2-prop-2-enylpiperidine	165.278	651
751	N-((1E)-2-phenyl-1-azaprop-1-enyl)-2-pyridylacetamide_chloride	254.311	652
752	N-pentylformamide	115.175	652
753	(cyclohexylamino)-N-(2,2,6,6-tetramethyl(4-piperidyl))carboxamide	281.44	653
754	1,2,5-trimethylpyrrole	109.171	653
755	2-[2-(4-oxo-3-hydroquinazolin-3-yl)acetylamino]benzamide	322.323	653
756	6-(azaperhydroepin-2-ylideneazamethyl)-2H,3H-benzo[e]1,4-dioxane	246.308	653
757	3-(2-aminophenyl)pentan-3-ol	179.261	654
758	4-[(2S,6S)-2,6-bis(2-methylprop-2-enyl)-4-1,2,5,6-tetrahydropyridyl]hepta-1,6-dien-4-ylamine	300.486	654
759	1-ethyl-3-[4-(2-oxo(3-hydrobenzimidazolyl))piperidyl]azolidine-2,5-dione	342.397	655
760	7-methyl-4-(piperazinylmethyl)chromen-2-one	258.319	655
761	diethyl[2-(2-methoxyphenoxy)ethyl]amine_ethanedioic_acid	223.314	655
762	(2,6-dioxo(1,3-dihydropyrimidin-4-yl))-N-(2-ethoxyphenyl)carboxamide	275.263	656
763	{[3-(dimethylamino)propyl]amino}-N-naphthylcarboxamide	271.361	656
764	1-methyl-4-piperidyl_2-phenylacetate	233.31	657
765	2-(1-cyclohexyl-2,5-dioxo(1,3-diazolidin-4-yl))-N-(3-pyridyl)acetamide	316.359	657
766	5-(1,5-dimethylpyrrol-2-yl)-1H-1,2,3,4-tetraazole	163.182	657
767	naphthyl-N-(3-pyridylmethyl)carboxamide	262.31	657
768	[(3-amino-1-phenyl(2-naphthyl))methyl]dimethylamine	276.38	658
769	{2-[(cyclohexylamino)methyl]phenyl}methan-1-ol	219.326	658
770	2-(2-methoxyphenyl)-2,5,6,7,8,8a-hexahydro-2,8a-diazaindolizine-1,3-dione	261.28	658
771	1,2,5-trimethyl-4-phenylpyridine	198.287	659
772	1,5-dimethyl-3-(2-oxo-2-piperidylethyl)-1,3,5-triazaperhydroine-2,4,6-trione	282.299	659
773	4-hydroxy-1-methylhydroquinolin-2-one	175.187	659
774	quinoline-7-carbaldehyde	157.171	659
775	(2-hydroxyphenyl)-N-methylcarboxamide	151.165	660
776	(4-methylphenyl)-N-{2-[(4-methylpiperazinyl)carbonyl]phenyl}carboxamide	337.421	660
777	1-(4-piperidylphenyl)azoline-2,5-dione	256.304	660
778	methyl_3-[((1R)-1-methyl-2-phenylethyl)amino]propanoate	221.299	660
779	2-benzylbenzimidazole	208.262	662
780	5,6-dimethyl-3-methylthio-1,2,4-triazine	155.217	662
781	3,3,9-trimethyl-2,3,4,5,6,7,8,9,4b,8a-decahydro-4b-azaphenanthrene-1,10-dione	261.363	663
782	4,5-dihydro-6H-1,3-thiazin-2-yl(4-methylphenyl)amine	206.305	663
783	6-propyl-2H-benzo[d]1,3-dioxolene-5-ylamine	179.218	663
784	(2-phenylethyl)bis(4-pyridylmethyl)amine	303.406	665
785	2,3-dimethyl-5,6,7-trihydroindol-4-one	163.219	665
786	2,6-di(2-pyridyl)pyridine	233.272	665
787	2-hydroxyphenyl_pyrrolidinyl_ketone	191.229	665
788	5,8-dimethoxy-2,4-dimethylquinoline	217.267	665
789	chromen-4-one	146.145	666
790	2,4,7-trimethylpyridino[2,3-b]quinoline-5-ylamine	237.304	667
791	amino-N-[(2-ethoxyphenyl)methyl]amide	194.233	667

792	N-cyclopentyl-2-(2-1,2,3,4-tetrahydroisoquinolyl)acetamide	258.363	667
793	3,3-dimethyl-1-prop-2-enyl-1,2,3,4-tetrahydroisoquinoline	201.311	668
794	5-(3-methylbutanoylamino)benzene-1,3-dicarboxylic_acid	265.265	668
795	(tert-butyl)quinazolin-4-ylamine	201.271	669
796	1-{[(2-phenylethyl)(hydroxyphosphoryl)]methyl}azaperhydroepin-2-one	295.318	669
797	2,2,4-trimethyl-1,2,3,4-tetrahydroquinolin-8-ol	191.272	669
798	2-amino-N-(2-naphthyl)propanamide_chloride	214.266	669
799	3-[(4-morpholin-4-ylphenyl)azamethylene]-2H-benzo[c]azolidinimine	306.366	669
800	(2-aminophenyl)-N-propylcarboxamide	178.233	670
801	2,3,3-trimethyl-3H-pyrrolo[3,2-h]quinoline	210.278	670
802	5,6,7,8,9-pentahydro-4aH-carbazole-3-ylamine	186.256	670
803	4-(1,3-dioxobenzo[c]azolidin-2-yl)-N-methyl-N-(1,2,2,6,6-pentamethyl(4-piperidyl))butanamide	399.532	671
804	4-isopropylaniline	135.208	671
805	5,6,7,8,10-pentahydroacridin-9-one	199.252	671
806	4-propylaniline	135.208	674
807	N-{(1E)-2-[4-(dimethylamino)phenyl]-1-azavinyl}(3-aminophenyl)carboxamide	282.344	674
808	4-[3-(1-carbamoyl-4-piperidyl)propyl]piperidinecarboxamide	296.412	675
809	5-(3-pyrazolylphenyl)-1H-1,2,3,4-tetraazole	212.213	675
810	N-[(1E)-2-(3-methylphenyl)-1-azavinyl]-2-pyridylacetamide_chloride	254.311	675
811	indazole	118.138	676
812	N-{2-[4-(dimethylamino)phenyl]ethyl}(cyclohexylamino)carboxamide	289.42	676
813	1-phenyl-1,2,3,4-tetraazole	146.151	677
814	3-(2H,3H-benzo[3,4-e]1,4-dioxan-6-ylazamethylene)isoindolylamine	279.298	677
815	3-[(4-ethylpiperazinyl)methyl]-2,5-dimethylbenzaldehyde	260.378	677
816	[4,6-bis(dimethylamino)(1,3,5-triazin-2-yl)]dimethylamine	210.281	678
817	2-methyl-4-(phenylmethylene)-1,3-oxazolin-5-one	187.198	678
818	3-(2-methylpyrimidin-4-yl)benzoic_acid	214.223	678
819	N-{4-[4-(2-methylpropanoyl)piperazinyl]phenyl}propanamide	303.403	678
820	4-hydroxy-5-phenylcyclopent-4-ene-1,3-dione	188.182	679
821	2,4,6-trimethylaniline	135.208	681
822	4-indol-3-ylbutanamide	202.255	681
823	1-(5-methyl-2-pyridyl)-1,2,3,4-tetraazole	161.166	683
824	2-cyclohexyl-1-methylbenzimidazole	214.31	684
825	tris(4-aminophenyl)methan-1-ol	305.379	684
826	(2-butylthioethyl)thiocarboxamidine	192.337	685
827	2-(2,4,7-trimethylindol-3-yl)ethylamine	202.299	685
828	3,4-dimethyl-1-(2-phenylethyl)pyridine	212.314	685
829	6-(1,3-dioxobenzo[c]azolidin-2-yl)-N-(2-pyridylmethyl)hexanamide	351.404	685
830	amino(4,8-dimethylquinazolin-2-yl)carboxamidine	215.257	685
831	2-(2-naphthyloxy)ethylamine	187.241	686
832	3-(phenylamino)cyclohex-2-en-1-one	187.241	687
833	3-acetylindole	159.187	687
834	4-[bis(4-hydroxyphenyl)methylene]cyclohexa-2,5-dien-1-one	290.318	687
835	5-(4-methylphenyl)-2H-1,2,3,4-tetraazole	160.178	687
836	5-imidazolyl-1-phenyl-1,2,3,4-tetraazole	212.213	687
837	8-methoxy-2,4-dimethyl-2,3-dihydrofurano[3,2-c]quinoline	229.278	687
838	[4-({(1E)-2-[4-(N,N-dimethylcarbamoyl)phenyl]-2-azavinyl}amino)phenyl]-N,N-dimethylcarboxamide	338.408	688
839	2-(diethylamino)-N-(2,4,6-trimethylphenyl)acetamide	248.367	688
840	naphthyl-1,3-thiazolin-2-ylamine	228.311	688
841	(6Z,4E)-7-amino-3,3,4-trimethyl-7-(4-methylphenyl)-5,6-diazahepta-4,6-dien-2-one_iodide	259.35	689

842	6-phenylphenanthridine-3,8-diamine	285.348	689
843	N-(6-methyl(2-pyridyl))quinoxalin-6-ylcarboxamide	264.286	689
844	N-[4-(carbamoylmethoxy)phenyl](3,4,5-trimethoxyphenyl)carboxamide	360.366	689
845	o-isopropylaniline	135.208	689
846	1-(2-phenylethyl)-1,4-diazaperhydroine-2,6-dione	218.255	690
847	1-naphthyl-5-oxopyrrolidine-3-carboxylic_acid	255.273	690
848	methyl_3-[N-benzylcarbamoyl]propanoate	221.255	690
849	2,7,8-trimethylquinolin-4-ol	187.241	691
850	2-amino-6-methylheptane	129.245	691
851	6-(phenylethyl)-1-benzyl-1,3,5,6,7,8-hexahydro-6,8-diazaquinazoline-2,4-dione	362.43	691
852	N-[(1E)-2-(2,4-dimethoxyphenyl)-1-azavinyl]aminoamide	223.231	691
853	(2E)-3-(2-quinolyl)prop-2-enoic_acid	199.209	692
854	3,6-bis(indol-3-ylmethyl)-1,4-diazaperhydroine-2,5-dione	372.426	692
855	butyl_3-(3,5-dioxo-2H,4H-1,2,4-triazin-6-yl)propanoate	241.246	692
856	4-hydroxy-3-propylhydroquinolin-2-one	203.24	693
857	methyl_3-[(3-imino-2H-benzo[c]azolidinylidene)azamethyl]benzoate	279.298	693
858	3-methyl-1-{4-[(3-methylphenyl)methyl]piperazinyl}butan-1-one	274.405	694
859	7-(4-hydroxyphenyl)-4-methyl-2-benzyl-3,4,6,7-tetrahydro-4,6-diazaisoindole-1,5-dione	349.388	694
860	[(3-amino-5-imino(1,2-diazolin-4-ylidene))azamethyl](4-methoxyphenyl)amine	232.244	695
861	1-(indol-3-ylmethyl)-3,5,5-trimethyl-2-pyrazoline	241.335	695
862	2-[(octan-2-yl)amino]ethan-1-ol	173.298	695
863	5,6-dimethylbenzotriazole	147.179	695
864	1-(4-methoxyphenyl)azoline-2,5-dione	203.197	696
865	2-(spiro[3,4-dihydroisoquinoline-3,1'-cyclohexane]ylamino)acetic_acid	272.346	696
866	3-oxo-5,6-diphenyl-2-hydropyridazine-4-carboxylic_acid	292.293	696
867	N-(3,4-dimethoxyphenyl)(3-oxo(2H,4H-benzo[3,4-e]1,4-oxazin-6-yl))carboxamide	328.324	696
868	7-methyl-2-oxohydroquinoline-3-carboxylic_acid	203.197	697
869	(4E)-hex-4-en-3-one	98.144	698
870	3-(dimethylamino)-1-(2-naphthyl)propan-1-one	227.305	698
871	6-amino-3-benzyl-3-hydroquinazolin-4-one	251.287	698
872	N'-(2-phenylethyl)-N-(2,2,6,6-tetramethyl(4-piperidyl))ethane-1,2-diamide	331.457	699
873	bis[4-(dimethylamino)phenyl]methan-1-ol	270.374	700
874	1-(4-methoxyphenyl)pyrrolidin-2-one	191.229	701
875	trimethyl[2-(5-methyl-5-phenyl(1,3-dioxolan-4-ylidene))ethyl]amine_iodide	248.344	701
876	4-aminophenyl_4-methylpiperidyl_ketone	218.298	704
877	5-methyl-1-[(phenylmethoxy)methyl]-1,3-dihydropyrimidine-2,4-dione	246.265	704
878	8-ethyl-2-methylquinolin-4-ol	187.241	704
879	2-propylaniline	135.208	705
880	3-({(1E)-2-[4-(dimethylamino)phenyl]-1-azavinyl}amino)-6-methyl-4H-1,2,4-triazin-5-one	272.309	706
881	diethyl(4-methyl(2-quinolyl))amine	214.31	706
882	N-allylaniline	133.193	706
883	N-cyclohexyl-2-{4-[2-(cyclohexylamino)acetyl]piperazinyl}acetamide	364.53	706
884	((1Z)-1-morpholin-4-yl-2-phenyl-2-azavinyl)phenylamine	281.357	707
885	(3-aminophenyl)phenylamine	184.24	707
886	2-pyridyl-N-[3-(2-(4-pyridyl)ethyl)phenyl]carboxamide	303.363	708
887	3-amino-5-(piperidylcarbonyl)phenyl_piperidyl_ketone	315.414	708
888	1-(2-hydroxyphenyl)-2,6-dimethyl-5-phenylhydropyrimidin-4-one	292.337	709
889	1-cyclohexylazoline-2,5-dione	179.218	709

890	3,4-diacetyl-1-ethyl-2,5-dimethylpyrrole	207.272	709
891	3-indolinylpropylamine	176.261	709
892	4-(1,2,3,4-tetrahydroquinolyl)-1,3,5-triazine-2-ylamine	227.268	709
893	2-methyl-1H,3H-naphtho[1,2-e]1,3-oxazine	199.252	710
894	4,7-dimethoxy-5-[4-methyl-3-(morpholin-4-ylmethyl)(4,5-dihydroisoxazol-5-yl)]-2H-benzo[d]1,3-dioxolene	364.397	710
895	2-[(2-methylindol-3-yl)methyl]benzimidazole	261.326	711
896	4-(phenylmethoxy)phenylamine	199.252	711
897	3,4,7,8-tetramethylpyridino[3,2-h]quinoline	236.316	712
898	(2S,6R)-2,6-bis(2-methylprop-2-enyl)-1,2,5,6-tetrahydropyridine	191.316	713
899	[2-(phenylamino)ethyl]benzylamine	226.321	713
900	3-(6-methyl-2-oxohydropyridyl)propanoic_acid	181.191	713
901	5-amino-3-phenyl-10-hydro-2-pyrrolino[1,2-a]quinazolin-2-one	275.309	713
902	indeno[3,2-e]pyridino[2,3-b]pyrazin-10-one	233.229	713
903	methyl_3-(2,4-dioxo-3,10-dihydropyrimidino[4,5-b]quinolin-10-yl)benzoate	347.329	713
904	N-(6-methyl(2-pyridyl))[(6-methyl(2-pyridyl))amino]carboxamide	242.28	713
905	indol-3-yl-N-propylcarboxamide	202.255	714
906	N-[2-(diethylamino)ethyl]{2-[(N-phenylcarbamoyl)amino]phenyl}carboxamide	354.451	714
907	N-phenyl-N'-{2-[4-benzylpiperazinyl]ethyl}ethane-1,2-diamide	366.462	714
908	4-methyl-2-(4-methylquinazolin-2-yl)quinazoline	286.335	715
909	N-[4-(dimethylamino)phenyl]-N'-(2-methylpropyl)ethane-1,2-diamide	263.339	715
910	3,3,5-trimethyl-1-methylthio-3,4-dihydroisoquinoline	219.344	716
911	4-(4-aminophenyl)piperazinyl_2-phenylphenyl_ketone	357.454	716
912	(phenylcyclohexyl)prop-2-enylamine	215.338	718
913	N-(4-methoxyphenyl){[4-(2-morpholin-4-yl-2-oxoethoxy)phenyl]amino}carboxamide	385.419	719
914	(2E)-4-indolinyl-4-oxobut-2-enoic_acid	217.224	720
915	(3-pyridylmethyl)-1,2,3-trihydrocyclopenta[2,1-b]quinolin-9-ylamine	275.352	720
916	7-methyl-5,6,7,8,10-pentahydroacridin-9-one	213.279	720
917	N-[(1E)-2-(4-ethylphenyl)-1-azavinyl]-2-pyridylacetamide_chloride	268.338	720
918	2-(1-cyclohexyl-2,5-dioxo(1,3-diazolidin-4-yl))-N-[3-(2-(4-pyridyl)ethyl)phenyl]acetamide	420.51	721
919	triisobutylamine	185.352	721
920	2-(2-imino-3-methyl(3-hydrobenzimidazolyl))-1-phenylethan-1-one	265.314	723
921	2-[(4,4-dimethyl-2,6-dioxocyclohexyl)-3-pyridylmethyl]-5,5-dimethylcyclohexane-1,3-dione	369.46	723
922	3-(dimethylamino)-1-[4-(methylethyl)phenyl]propan-1-one	219.326	724
923	5-[((1E)-2-indol-3-yl-1-azavinyl)amino]-6-methyl-2H-1,2,4-triazin-3-one	268.277	724
924	5-hydroxy-2-azabicyclo[4.4.0]dec-2-yl_phenyl_ketone	259.347	724
925	N-(4-methylphenyl)-N'-{2-[4-(phenylcarbonyl)piperazinyl]ethyl}ethane-1,2-diamide	394.472	724
926	(1S,10S,11S,15S,2R,14R)-14-hydroxy-14-(2-hydroxyacetyl)-2,15-dimethyltetracyclo[8.7.0.0<2,7>.0<11,15>]heptadeca-3,6-diene-5,17-dione	358.433	725
927	2-(3-pyridyl)indole	194.235	725
928	indole-4-carbaldehyde	145.16	725
929	3-(3-methylbutanoylamino)benzoic_acid	221.255	726
930	4-(2-pyridylmethylthio)quinazoline	253.321	726
931	(2-aminophenyl)-N-(cyclopentylideneazamethyl)carboxamide	217.27	727
932	1-methylindole-3-carboxylic_acid	175.187	727
933	2,3,4-trihydrobeta-carbolin-1-one	186.213	727
934	(2-iminochromen-3-yl)-N-(4-methoxyphenyl)carboxamide	294.309	728

935	5-methyl-2-(1-methyl(4-pyridyl))indole	223.297	728
936	(4,6-dimethylpyrimidin-2-yl)[imino(5-phenyl(2-pyrazolinyl))methyl]amine	294.358	729
937	1-(3,4-dimethylphenyl)-5-oxopyrrolidine-3-carboxylic_acid	233.266	729
938	6-methyl-2-(4-pyridyl)indole	208.262	729
939	(4-phenylphenyl)-N-(4-pyridylmethyl)carboxamide	288.348	730
940	N-(4-{N-[(1E)-2-(3,4-dihydroxyphenyl)-1-azavinyl]carbamoyl}phenyl)benzamide	375.383	730
941	(3,4-dimethylphenyl)(5-methyl(1,3-thiazolin-2-yl))amine	220.332	731
942	morpholin-4-yl-N-(4-{[4-(morpholin-4-ylcarbonylamino)phenyl]methyl}phenyl)carboxamide	424.499	731
943	N-((1E)-2-quinoxalin-2-yl-1-azavinyl)-2-(4-oxo(3-hydroquinazolin-3-yl))acetamide	358.359	731
944	[(1E)-2-(4-methylphenyl)-1-azavinyl](4,6-dimethylpyrimidin-2-yl)amine	240.307	732
945	1,7-dimethylbenzo[d]azolidine-2,3-dione	175.187	732
946	1-benzyl-3-[4-benzylpiperazinyl]azolidine-2,5-dione	363.458	732
947	ethyl_(2E)-3-(N-{3-[N-(2,4-dimethylphenyl)carbamoyl]propanoylamino}carbamoyl)prop-2-enoate	361.397	732
948	N-[(1E)-2-(1-methylpyrrol-2-yl)-1-azavinyl]-N'-[(1E)-2-(1-methylpyrrol-2-yl)-1-azavinyl]pentane-1,5-diamide	342.4	732
949	2-(3,3-dimethyl-2,3,4-trihydroisoquinolylidene)cyclohexane-1,3-dione	269.343	733
950	1-(phenylmethylthio)-2-(4-pyridyl)ethane	229.339	734
951	1-ethyl-2,5-dimethylpyrrole-3-carbaldehyde	151.208	735
952	2-acetyl-1,2,3,4-tetrahydropyridino[4,3-b]indole	214.266	736
953	3-[3-(1,3-dioxobenzo[c]azolidin-2-yl)propanoylamino]benzoic_acid	338.319	736
954	4-((1E)-4-cyclohex-1-enyl-2-azabut-1-enyl)-1,2-dimethoxybenzene	273.374	736
955	di-3-amino-4-methylphenyl_ketone	240.304	736
956	methyl_4-cyclohexyl-5-oxo-1,2,4-triazoline-3-carboxylate	225.247	738
957	N-(3,4-dimethoxyphenyl)[(3,4-dimethoxyphenyl)amino]carboxamide	332.355	738
958	5-methylindole-3-carbaldehyde	159.187	739
959	2-(1,3-diethyl-2,5-dioxo(1,3-diazolidin-4-yl))-N-[2-(4-methoxyphenyl)ethyl]acetamide	347.413	741
960	3-phenyl-N-[4-(4-pyridylmethyl)phenyl]propanamide	316.402	741
961	N-[(1E)-2-(4-methylphenyl)-1-azavinyl](4-aminophenyl)carboxamide	253.303	741
962	3-(3-aminophenyl)chromen-2-one	237.257	743
963	[4-(2,5-dioxoazolidinyl)phenyl]-N-(4-methoxyphenyl)carboxamide	324.335	744
964	diethyl[2-(naphthylamino)ethyl]amine	242.363	744
965	[(4,6-dimethylpyrazolo[5,4-b]pyridin-3-yl)diazenyl]diethylamine	246.314	745
966	diethyl_2-{amino[(2-methylpropanoylamino)amino]methylene}propane-1,3-dioate	287.315	745
967	(2-imidazolin-2-ylmethyl)phenylbenzylamine	265.357	746
968	3-[(dimethylamino)amino]-5-phenylcyclohex-2-en-1-one	230.309	746
969	2-{[4-benzylpiperidyl]methyl}pyridine	266.385	747
970	N-[2-(4-methylphenoxy)ethyl]acetamide	193.245	747
971	(2E)-3-(4-aminophenyl)-2-phenylprop-2-enoic_acid	239.273	748
972	1,2-bis(2-pyridylmethyl)-4-phenyl-1,2,4-triazolidine-3,5-dione	359.387	748
973	4-pyridyl-N-(2,4,6-trimethylphenyl)carboxamide	240.304	748
974	methyl_3-(4-aminophenyl)-2-phenylpropanoate	255.316	749
975	4-[6-(4-hydroxyphenyl)-2-pyridyl]phenol	263.295	750
976	5-(6-methyl-2-quinolyl)-2H-1,2,3,4-tetraazole	211.226	750
977	diethyl(2-(4-pyridyl)quinazolin-4-yl)amine	278.356	750
978	N-(2,5-dimethylphenyl)-N'-(2-hydroxypropyl)ethane-1,2-diamide	250.297	750
979	N-[4-(dimethylamino)phenyl][1-(4-ethylphenyl)-5-oxopyrrolidin-3-yl]carboxamide	351.447	750
980	4-(dimethylamino)phenyl_pyrrolidinyl_ketone	218.298	751

981	methylene_blue	284.398	752
982	(2,3-dimethyl-1,4-dioxo(2,3-dihydrophthalazin-6-yl))-N-(2-quinolyl)carboxamide	360.371	753
983	(2-methyl(4-quinolyl))(3-methylphenyl)amine	248.327	753
984	2-[5-methyl-2-(methylethyl)phenoxy]ethylamine	193.288	753
985	4-(phenylazamethylene)-5-hydro-1,3,5-triazino[6,1-b]benzothiazole-2-ylamine	293.345	753
986	1,3-diethyl-2-oxo-3-hydrobenzimidazole-5-carbaldehyde	218.255	754
987	11H-dibenzo[b,f]1,4-dioxepane-2-ylamine	213.235	754
988	(phenylamino)-N-(2-pyridyl)carboxamide	213.238	755
989	2-(2H,3H-benzo[e]1,4-dioxin-2-yl)-1-morpholin-4-ylethan-1-one	263.293	755
990	N-[1-benzyl(4-piperidyl)]-2-propylpentanamide	316.486	755
991	4-(2-hydroxy-2,2-diphenylethyl)-1,2,5-trimethylpiperidin-4-ol	339.477	756
992	5-(2-phenyl-1,2,3-triazol-4-yl)-1H-1,2,3,4-tetraazole	213.201	756
993	N-{4-[3,4-di(propanoylamino)phenoxy]phenyl}propanamide	383.446	756
994	((1E)-1-amino-5-methyl-2-azahex-1-enyl)(4,6-dimethylpyrimidin-2-yl)amine	235.331	757
995	[(9-ethylcarbazol-3-yl)methyl]methylamine	238.332	757
996	3-[4-(acetylamino)phenyl]-2-phenylpropanoic_acid	283.326	757
997	1-(4-ethylphenyl)-5-oxopyrrolidine-3-carboxylic_acid	233.266	758
998	2,2-dimethyl-N-(phenylamino)propanamide	192.26	758
999	4-{[3-(butanoylamino)phenyl]carbonylamino}benzoic_acid	326.351	758
1000	5-(azaethylidene)-2,3-diphenyl-1,2,4-thiadiazoline_chloride	267.348	758
1001	3,4,5-trimethylpyrrole-2-carbaldehyde	137.181	759
1002	6-(diphenylamino)-4-morpholin-4-yl-1H-1,3,5-triazin-2-one	349.391	759
1003	cymarine	548.672	759
1004	2-({4-[(3-methylphenyl)methyl]piperazinyl}methyl)benzo[c]azoline-1,3-dione	349.432	760
1005	3,7-dimethylindeno[2,3-c]pyridin-9-one	209.247	760
1006	4-phenylmorpholine	163.219	760
1007	6-cyclohexyl-1,4,5,7-tetramethylpyrrolo[3,4-d]pyridazine	257.378	760
1008	N-[4-(5-phenyl-2-pyrazolin-3-yl)phenyl]acetamide	279.341	760
1009	1,2,3-trimethylcyclohepta[1,2-c]pyrrol-6-one	187.241	761
1010	2-(2-imino-3-methyl(3-hydrobenzimidazolyl))-1-(4-methylphenyl)ethan-1-one	279.341	761
1011	5-methyl-2-(2,4,5-trimethylphenyl)pyridine	211.306	761
1012	dibenzo[c,e]1,2-dithiane-3,8-diamine	246.344	764
1013	(1-methyl-5-oxopyrrolidin-3-yl)-N-[3-(phenylcarbonyl)phenyl]carboxamide	322.363	765
1014	N-{(1E)-2-[4-(methylethyl)phenyl]-1-azavinyl}-2-pyridylacetamide_chloride	282.365	765
1015	(3Z)-4-amino-4-phenylbut-3-en-2-one	161.203	766
1016	2-naphthyl-1-[4-benzylpiperazinyl]ethan-1-one	344.455	766
1017	(2S)-2-amino-N-(2-naphthyl)-3-phenylpropanamide	290.364	768
1018	4-(2,5-dimethylphenyl)-2,5-dimethylpyridine	211.306	768
1019	methyl_2-{1-[(methoxycarbonyl)methyl]-2,4-dioxo-1,3-dihydroquinazolin-3-yl}acetate	306.274	771
1020	methyl_quinoline-2-carboxylate	187.198	771
1021	(4,6-dimethylpyrimidin-2-yl)[imino(naphthylamino)methyl]amine	291.355	772
1022	1-(4-methoxyphenyl)-3-oxolan-2-ylazolidine-2,5-dione	275.304	772
1023	2-(2,4-dimethylphenyl)-1-methylquinoxaline	249.335	772
1024	(2E)-3-[N-(2,4-dimethylphenyl)carbamoyl]prop-2-enoic_acid	219.24	774
1025	4-methyl-2-phenyl-3-pyrazolin-5-one	174.202	774
1026	(2-ethylhexyl)(2-pyridylmethyl)amine_chloride_chloride	220.357	775
1027	(4,6-dimethylpyrimidin-2-yl)(2-methylphenyl)amine	213.282	775

1028	5-[(3,4-dimethoxyphenyl)methyl]-1-(4-methoxyphenyl)-1,3,5-trihydropyrimidine-2,4,6-trione	384.388	775
1029	N'-ethyl-N-(2-indol-3-ylethyl)ethane-1,2-diamide	259.307	775
1030	1-{2-[4-(methylethyl)phenoxy]acetyl}piperidine-4-carboxamide	304.388	776
1031	ethyl_2-[3-(ethoxycarbonyl)-2-pyridyl]pyridine-3-carboxylate	300.313	776
1032	N-(phenylethyl)[4-(2-1,2,3,4-tetrahydroisoquinolylmethyl)phenyl]carboxamide	370.493	776
1033	N-[4-(2-oxopyrrolidinyl)phenyl](3,4,5-trimethoxyphenyl)carboxamide	370.404	776
1034	N-{(1E)-2-[4-(diethylamino)phenyl]-1-azavinyl}benzamide	295.383	776
1035	(4-methoxyphenyl)(phenylethyl)amine	227.305	777
1036	1,3-dibutylurea	172.27	777
1037	N-(2,5-dimethoxyphenyl)-2-(3,5-dimethyl-2,4,6-trioxo(1,3,5-triazaperhydroinyl))acetamide	350.33	777
1038	2-butoxyphenylamine	165.235	778
1039	3-acetyl-2-oxochromene	188.182	778
1040	ethyl_(3E)-4-{[4-(acetylamino)phenyl]carbonylamino}-3-methyl-2-(methylethyl)-4-azabut-3-enoate	347.413	778
1041	methyl_4-[(propylamino)carbonylamino]benzoate	236.27	778
1042	N-[4-(4-oxo-3-phenyl-1,2,3-trihydroquinazolin-2-yl)phenyl]acetamide	357.411	780
1043	2-[6-((4aS,9bR)-2,8-dimethylpiperidino[4,3-b]indolin-5-yl)-6-oxohexyl]benzo[c]azoline-1,3-dione	445.56	781
1044	N-((1E)-2-indol-2-yl-1-azavinyl)benzamide	263.298	781
1045	5-[(1Z)-1-methyl-2-(phenylamino)-2-azavinyl]-4-methyl-6-phenyl-1,3,6-trihydropyrimidin-2-one	320.393	782
1046	9,10-dimethoxy-5,6-dihydro-2H-1,3-dioxoleno[4,5-g]isoquinolino[3,2-a]isoquinoline	336.366	783
1047	1,2-bis(2-methylpropyl)benzimidazole	230.352	785
1048	5-{[(4-hydroxy-2-methylphenyl)amino]methylene}-1-(4-methylphenyl)-1,3-dihydropyrimidine-2,4,6-trione	351.361	785
1049	benzyldi(2-pyridyl)amine	261.326	786
1050	4,5-diphenyl-2-(3-pyridyl)imidazole	297.359	787
1051	(2Z)-3-(dimethylamino)-1-(3-methylphenyl)prop-2-en-1-one	189.257	788
1052	4-aminophenyl_3-phenylpyrrolidinyl_ketone	266.342	788
1053	(4-aminophenyl)-N-(4-methylphenyl)carboxamide	226.277	789
1054	1-(2-methylphenyl)-5-[(1-methylpyrrol-2-yl)methylene]-1,3-dihydropyrimidine-2,4,6-trione	309.324	790
1055	2,3,4-trimethylbenzo[h]quinoline	221.301	791
1056	2-phenyl-3-hydroquinazolin-4-one	222.246	792
1057	4-indol-3-ylbutanoic_acid	203.24	792
1058	benzyl(hexyl)dimethylazanium_chloride	220.377	792
1059	1-cyclohexyl-2-pyrrolino[2,3-b]quinoline-4-ylamine	267.373	793
1060	ethyl_2-{N-[4-(methoxycarbonyl)phenyl]carbamoyl}acetate	265.265	793
1061	(2-aminophenyl)(2-phenoxyethyl)amine	228.293	794
1062	2-aminophenyl_2-methylpiperidyl_ketone	218.298	794
1063	[2-(5-hexyl-5-methyl(1,3-dioxolan-4-ylidene))ethyl]trimethylamine_iodide	256.408	795
1064	1-methyl-4-phenoxyphthalazine	236.273	795
1065	3-[N-(1,3-dioxobenzo[c]azolin-2-yl)carbamoyl]phenyl_acetate	324.292	796
1066	N-(4-{N-[2-(5-methylindol-3-yl)ethyl]carbamoyl}phenyl)acetamide	335.405	796
1067	(4S)-2-{1-[(4S)-4-(tert-butyl)(1,3-oxazolin-2-yl)]-isopropyl}-4-(tert-butyl)-1,3-oxazoline	294.436	797
1068	cyclopropyl[(9-ethylcarbazol-3-yl)methyl]amine	264.369	797
1069	1-(indol-3-ylmethyl)-4-[(2-methylphenyl)methyl]piperazine	319.449	798
1070	1-cyclohexyloxy-2-[4-(2-cyclohexyloxyethyl)piperazinyl]ethane	338.532	798

1071	2,6,10-trimethylpyrazolo[1,5-c]pyrazolo[1,5-e]pyrazolo[1,5-a]1,3,5-triazaperhydroine	240.267	798
1072	6-methyl-1-{[2-(4-methylphenoxy)ethoxy]methyl}-1,3-dihydropyrimidine-2,4-dione	290.318	798
1073	3-[N-(diphenylmethyl)carbamoyl]propanoic_acid	283.326	800
1074	N-[(4-hydroxyphenyl)ethyl](4-methylphenyl)carboxamide	255.316	800
1075	(2E)-3-amino-3-[(3-methylquinoxalin-2-yl)amino]-1-phenyl-2-azaprop-2-en-1-one	305.338	801
1076	(4-aminophenyl)-N-naphthylcarboxamide	262.31	801
1077	N-(1-carbamoyl-3-methylbutyl)(2-methoxy-3-methylphenyl)carboxamide	278.35	801
1078	3-oxobenzo[f]chromene-2-carboxamide	239.23	802
1079	2-(phenylcarbonyl)-1,2,3,4-tetrahydroisoquinoline-3-carboxylic_acid	281.31	804
1080	3-(N-(2-5,6,7,8-tetrahydronaphthyl)carbamoyl)propanoic_acid	247.293	804
1081	N-((1E)-2-(3-pyridyl)-1-azavinyl)[4-(tert-butyl)phenyl]carboxamide	281.357	804
1082	N-(4-aminophenyl)[4-(tert-butyl)phenyl]carboxamide	268.358	804
1083	{(1E)-1-methyl-2-[4-(phenyldiazenyl)phenyl]-2-azavinyl}dimethylamine_chloride	266.345	805
1084	3,3,6,6,9-pentamethyl-2,3,4,5,6,7,9,10-octahydroacridine-1,8-dione	287.401	805
1085	methyl_indole-3-carboxylate	175.187	805
1086	N-[4-({[N-(2,4-dimethylphenyl)carbamoyl]methyl}amino)phenyl]propanamide	325.41	805
1087	1,3-dimethyl-5-[(2,3,4-trimethoxyphenyl)methyl]-1,3,5-trihydropyrimidine-2,4,6-trione	336.344	806
1088	2-methyl-3-oxo-5,6-diphenyl-2-hydropyridazine-4-carboxylic_acid	306.32	806
1089	3-(phenylcarbonyl)hydroquinoxalin-2-one	250.256	806
1090	2-[N-(tert-butyl)carbamoyl]phenyl_acetate	235.282	807
1091	5-(phenylcarbonyl)-5-azabicyclo[4.4.0]decan-2-one	257.332	807
1092	2-[4-(4-methylphenyl)-1-oxo(2-hydrophthalazin-2-yl)]-N-(3-pyridyl)acetamide	370.41	809
1093	methyl_3-[N-(2,2-diphenylacetylamino)carbamoyl]propanoate	340.378	809
1094	N-[4-(diethylamino)phenyl]-2-ethylhexanamide	290.448	810
1095	(4,6-diimidazolyl(1,3,5-triazin-2-yl))methylphenylamine	318.34	811
1096	ethyl_quinoxaline-2-carboxylate	202.212	811
1097	[4-(dimethylamino)phenyl]-N-(3-ethyl-4-oxo(3-hydroquinazolin-7-yl))carboxamide	336.393	812
1098	6-amino-2-phenylbenzo[d]1,3-oxazin-4-one	238.245	812
1099	1-pentyl-2-propylbenzimidazole	230.352	813
1100	8-methylthioquinoline	175.248	813
1101	N-(1,4-dioxo-3-pyrrolidinyl(2-naphthyl))-N-methylacetamide	298.341	813
1102	N-(diphenylmethyl)-N'-(2-hydroxyethyl)ethane-1,2-diamide	298.341	813
1103	(phenylamino)-N-[(3,4,5-trimethoxyphenyl)methyl]carboxamide	316.356	814
1104	[6-(dimethylamino)-4-piperidyl(1,3,5-triazin-2-yl)]dimethylamine	250.346	814
1105	3-(diethylamino)-1-cyclohexyl-1-phenylpropan-1-ol	289.46	815
1106	4-(tert-butyl)phenyl_4-benzylpiperazinyl_ketone	336.476	815
1107	2,5-dimethyl-1-(methylethyl)pyrrole-3-carbaldehyde	165.235	816
1108	2-methylpiperidyl_5-[(2-methylpiperidyl)carbonyl](3-pyridyl)_ketone	329.441	816
1109	1-(2-methylpropyl)-2-benzylbenzimidazole	264.369	817
1110	N,N-dimethylnonylamine	171.325	817
1111	5,6-diphenylpyrazin-2-ol	248.284	818
1112	ethyl_2-(3-methyl-2-oxo-1,3,4-trihydroquinazolinyl)acetate	248.281	818
1113	4,6-dimethoxyindole	177.202	821
1114	N-[4-(4-pyridylmethyl)phenyl]-2-quinolylcarboxamide	339.396	821
1115	1-methylindoline-5-carbaldehyde	161.203	822
1116	indeno[3,2-b]pyridin-5-one	181.193	822

1117	(2E)-N-(6-methyl(2-pyridyl))-3-phenylprop-2-enamide	238.288	823
1118	1-indol-3-yl-10-hydro-1,2,4-triazolo[4,3-a]quinoline	284.32	823
1119	3-{[(1E)-2-(2-hydroxyphenyl)-1-azavinyl]amino}-6-benzyl-2H-1,2,4-triazin-5-one	321.338	823
1120	7,8-dimethoxy-2-phenyl-2-hydrophthalazin-1-one	282.298	823
1121	2-[(1-propylindol-3-yl)methyl]propanedioic_acid	275.304	826
1122	3-[4-(diphenylmethyl)piperazinyl]-1-phenylazolidine-2,5-dione	425.529	826
1123	N-cyclopentyl-2-[7-methyl-2,4-dioxo-1-phenyl-5-(pyrrolidinylcarbonyl)(1,3-dihydropyridino[2,3-d]pyrimidin-3-yl)]acetamide	475.546	826
1124	((1Z)-1-phenyl-2-pyrazolylvinyl)pyrazole	236.276	828
1125	6-methyl-2-(6-methyl(3-quinolyl))-5-quinolylamine	299.374	828
1126	(2S,5S,6S,9S)-5,9,13-trimethyl-3-oxatricyclo[7.4.0.0<2,6>]trideca-1(13),10-diene-4,12-dione	246.305	829
1127	methyl_2-(methoxycarbonyl)indole-3-carboxylate	233.223	829
1128	1-methylindole-5-carbaldehyde	159.187	830
1129	(3,4-dimethylphenyl)-N-(2-pyridyl)carboxamide	226.277	831
1130	2-methoxy-5-phenylphenylamine	199.252	831
1131	3-(1,2,3,4-tetrahydroquinolyl)propanoic_acid	205.256	831
1132	ethyl_4-(pyrrolidinylcarbonylamino)benzoate	262.308	831
1133	N-((1E)-2-naphthyl-1-azavinyl)(4-aminophenyl)carboxamide	289.336	831
1134	phenyl(prop-2-enylcyclohexyl)amine	215.338	831
1135	[(4-methoxyphenyl)methyl]{[4-(methylethyl)phenyl]methyl}amine	269.386	834
1136	2-(1,3-dioxobenzo[c]azolidin-2-yl)-3-methylbutanoic_acid	247.25	834
1137	4-({5-(2,2-dimethylpropanoylamino)-3-[N-(4-carboxyphenyl)carbamoyl]phenyl}carbonylamino)benzoic_acid	503.51	835
1138	4-quinoxalin-2-ylphenol	222.246	835
1139	N-[(1E)-2-(2,4-dihydroxy-6-methylphenyl)-1-azavinyl](2-hydroxyphenyl)carboxamide	286.287	835
1140	{3-[(1,3-dimethyl-2,4-dioxo(1,3-dihydroquinazolin-6-yl))carbonylamino]-4,5-dimethoxyphenyl}-N,N-diethylcarboxamide	468.508	836
1141	5-[(2,4-dimethoxyphenyl)methyl]-1-phenyl-1,3,5-trihydropyrimidine-2,4,6-trione	354.362	836
1142	[4-(diethylamino)phenyl]-N-(2-methylpropyl)carboxamide	248.367	837
1143	N,N-bis(indol-4-ylmethyl)acetamide	317.39	838
1144	((1E)-2-(3-pyridyl)-1-azavinyl)(4,6-dipiperidyl(1,3,5-triazin-2-yl))amine	366.468	839
1145	6-(diphenylamino)-4-ethoxy-1H-1,3,5-triazin-2-one	308.339	839
1146	(1S,5S)-4,6,6-trimethylbicyclo[3.1.1]hept-3-en-2-one	150.22	840
1147	1,2,3,4,9-pentahydro-4aH-carbazol-1-ol	187.241	840
1148	3-[6-(1,3-dioxobenzo[c]azolidin-2-yl)hexanoylamino]benzoic_acid	380.399	840
1149	4-methyl-2-benzylspiro[2,4,5,6,3a,6a-hexahydro-2,5-diazapentalene-6,3'-indoline]-1,3,10-trione	361.399	840
1150	N-[(1E)-2-(2,3-dimethoxyphenyl)-1-azavinyl]benzamide	284.314	840
1151	(1Z)-2-(4,4-dimethyl(1,3-oxazolin-2-yl))-1-(4-methoxyphenyl)vinylamine	246.308	841
1152	1-methylindole-2-carboxylic_acid	175.187	841
1153	4-{[4-(diphenylmethyl)piperazinyl]carbonyl}-2-hydrophthalazin-1-one	424.501	841
1154	4-pyrrolylphenol	159.187	842
1155	2-{(1E)-2-[4-(diethylamino)phenyl]-2-azavinyl}-5-(diethylamino)phenol	339.48	843
1156	6-pyrrolyl-1H-indazole	183.212	843
1157	methyl_3-[N-(2,4-dimethylphenyl)carbamoyl]propanoate	235.282	843
1158	{4-[(1E)-1-methyl-2-(phenylamino)-2-azavinyl]-2-methyl-6-oxo-1-(2-pyridyl)(3-hydropyridyl)}-N-(2-pyridyl)carboxamide	438.488	845
1159	3-[((1E)-2-indol-3-yl-1-azavinyl)amino]-6-phenyl-4H-1,2,4-triazin-5-one	330.348	845
1160	3-[1-(carboxymethyl)-5-(4-methylphenyl)pyrrol-2-yl]propanoic_acid	287.315	845

1161	N-[2-(3,4-dimethoxyphenyl)ethyl](1,3-dimethyl-2,6-dioxo(1,3-dihydropyrimidin-4-yl))carboxamide	347.37	845
1162	2-((1E)-2-phenyl-2-azavinyl)-5-(diethylamino)phenol	268.358	846
1163	10-methyl-10-hydroacridin-9-one	209.247	848
1164	2-(2-oxo-3-phenylhydroquinoxalinyl)acetic_acid	280.282	848
1165	2-phenylquinoline-4-carboxylic_acid	249.268	848
1166	N,N-bis(methylethyl)[4-(N-(2-pyridyl)carbamoyl)phenyl]carboxamide	325.41	848
1167	N-[(4-hydroxyphenyl)ethyl]naphthylcarboxamide	291.349	848
1168	(1S,2R)-2-[N-(2,2-diphenylacetylamino)carbamoyl]cyclohexanecarboxylic_acid	380.443	849
1169	2-acetyl-3,5-dimethyl-4-(phenylcarbonyl)pyrrole	241.289	850
1170	3,5-bis(tert-butyl)phenyl_4-methylpiperazinyl_ketone	316.486	850
1171	(4-methylphenyl)-N-(2-pyrrolylethyl)carboxamide	228.293	851
1172	N-[4-hydroxy-2-methyl-5-(methylethyl)phenyl]-2-(4-methyl-1-oxo(2-hydrophthalazin-2-yl))acetamide	365.431	852
1173	3-(2,3,3-trimethylindolinyl)propanamide	232.325	853
1174	dihexylamine	185.352	853
1175	N-phenyl(2,4,5-trimethylpyrrol-3-yl)carboxamide	228.293	853
1176	5,6,7,8,9-pentahydro-4aH-carbazole-3-carboxylic_acid	215.251	854
1177	indol-3-yl-N-(2-indol-3-ylethyl)carboxamide	303.363	854
1178	1,1,3,3-tetraethylurea	172.27	855
1179	2-[(3-methylquinoxalin-2-yl)methyl]-4,7,3a,7a-tetrahydroisoindole-1,3-dione	307.351	856
1180	4-piperidyl-2-pyrrolidinylquinazoline	282.388	856
1181	8-ethylthioquinoline	189.275	856
1182	bis[2-(phenylmethoxy)ethyl]dimethylamine	314.447	856
1183	methyl_4-[2-(4-methyl-1-oxo-2-hydrophthalazin-2-yl)acetylamino]benzoate	351.361	856
1184	2-methyl-3-(3-methylphenyl)-3-hydroquinazolin-4-one	250.299	857
1185	6-{[4-(diphenylmethyl)piperazinyl]carbonyl}-1,3-dimethyl-1,3-dihydropyrimidine-2,4-dione	418.494	857
1186	8-(N-acetylacetylamino)-2-naphthyl_acetate	285.299	857
1187	2-(2-(5,6,7,8,9-pentahydro-4aH-carbazol-9-yl)ethyl)benzimidazole	315.417	858
1188	2,15-dimethyltetracyclo[8.7.0.0<2,7>.0<11,15>]heptadec-6-ene-5,14,17-trione	300.397	858
1189	8-(4-methoxyphenyl)-5,7,8-trihydro-2H-1,3-dioxolano[4,5-g]quinolin-6-one	297.31	858
1190	1-phenyl-2-pyrrolylethan-1-ol	187.241	859
1191	2-(1-ethylindol-3-yl)ethan-1-ol	189.257	859
1192	3-cyclohexyl-N-(4-oxo(3-hydroquinazolin-6-yl))propanamide	299.372	859
1193	bis(2-cyclohexyloxyethyl)amine	269.426	859
1194	phenyl(2-phenylquinazolin-4-yl)amine	297.359	859
1195	phenylmethyl_4-carbamoyl-3-phenylbutanoate	297.353	859
1196	(3E)-4-(4-methoxyphenyl)but-3-en-2-one	176.215	861
1197	5-pyrrolyl-3-hydroisobenzofuran-1-one	199.209	861
1198	(2-{N-[2-(2-methylindol-3-yl)ethyl]carbamoyl}pyrrolidinyl)-N-benzamide	390.484	862
1199	(3E)-4-phenylbut-3-en-2-one	146.188	862
1200	2-(2-methylphenyl)-1,3-dioxobenzo[c]azoline-5-carboxylic_acid	281.267	862
1201	2-(4-phenyl-3-1,4-dihydroquinolyl)quinoline	334.42	862
1202	1-acetyl-3-(1,3-dioxo(4,5,6,7,3a,7a-hexahydroisoindol-2-yl))benzene	271.315	863
1203	4-(4-indol-3-ylbutanoyl)-1,3,4-trihydroquinoxalin-2-one	333.389	863
1204	(2E)-3-(3,4-dimethoxyphenyl)-N-[2-(2-methoxyphenoxy)ethyl]prop-2-enamide	357.405	864

1205	2-[5-(N,N-diethylcarbamoyl)-1,7-dimethyl-2,4-dioxo(1,3-dihydropyridino[2,3-d]pyrimidin-3-yl)]-N-phenylacetamide	423.471	864
1206	indolo[2,3-b]quinoxaline	219.245	864
1207	2,6-dimethyl-4-(3-methylphenoxy)quinoline	263.338	865
1208	4-[3-(1,3-dimethylpyrazol-4-yl)-1-(4-methylphenyl)pyrazol-5-yl]-1,3-dimethylpyrazole	346.434	865
1209	2-[2-(8-methoxy-2-1,2,3,4-tetrahydroquinolyl)ethyl]phenylamine	282.385	866
1210	4-(N-methyl-N-phenylcarbamoyl)phenyl_acetate	269.299	866
1211	coproporphyrin	654.718	866
1212	N-(2H,3H-benzo[3,4-e]1,4-dioxin-6-yl)(phenylamino)carboxamide	270.287	866
1213	N-[4-(tert-butyl)cyclohexyl](7-methyl-2,4-dioxo(1,3-dihydropyridino[2,3-d]pyrimidin-5-yl))carboxamide	358.439	866
1214	N-(diphenylmethyl)-3-pyridylcarboxamide	288.348	867
1215	2-(1-cyclohexyl-2,5-dioxo(1,3-diazolidin-4-yl))-N-(2-indol-3-ylethyl)acetamide	382.461	868
1216	2,6-bis(2-methylprop-2-enyl)-1-benzyl-1,2,5,6-tetrahydropyridine	281.44	868
1217	8-(2,3-dimethoxyphenyl)-5,7,8-trihydro-2H-1,3-dioxolano[4,5-g]quinolin-6-one	327.336	868
1218	phenyl_pyridine-3-carboxylate	199.209	868
1219	(1E,3E)-1,4-bis(2-methylindol-3-yl)-2,3-diazabuta-1,3-diene	314.389	870
1220	3-(5-phenylpyrrol-2-yl)propanoic_acid	215.251	870
1221	6-methyl-2,4-diphenylpyridine	245.323	870
1222	(2E)-1-(2,4-dihydroxyphenyl)-3-(2-hydroxyphenyl)prop-2-en-1-one	256.257	872
1223	2-phenyl-4-piperidylquinoline	288.391	872
1224	1-(2-phenoxyethyl)-3-hydrobenzimidazol-2-one	254.288	874
1225	1-(3,5-dimethylphenyl)-1,2,3,4-tetraazole	174.205	875
1226	1,2,3-trihydrobenzo[g]quinolin-4-one	197.236	875
1227	3-[3-(3-aminophenoxy)phenoxy]phenylamine	292.337	876
1228	N-[(1E)-2-(3-hydroxyphenyl)-1-azavinyl](3-hydroxy(2-naphthyl))carboxamide	306.32	876
1229	N-methyldihexylamine	199.379	876
1230	(2E)-1-(2-hydroxyphenyl)-3-(3-hydroxyphenyl)prop-2-en-1-one	240.258	877
1231	N-[2-(4-methyl-2-quinolyl)phenyl]acetamide	276.337	877
1232	(2E)-3-(dimethylamino)-1-(2-naphthyl)prop-2-en-1-one	225.29	879
1233	1-methyl-2-(3-(2-quinolyl)(4-quinolyl))benzimidazole	386.455	879
1234	N,N-dimethyldecylamine	185.352	879
1235	4-(phenylamino)chromen-2-one	237.257	880
1236	5-(4-methylphenyl)pyrrole-2-carboxylic_acid	201.224	880
1237	8-methoxy-2,2,4-trimethyl-1,2,3,4-tetrahydroquinoline	205.299	880
1238	methylethyl_6-cyclohex-3-enyl-4-methyl-2-oxo-1,3,6-trihydropyrimidine-5-carboxylate	278.35	880
1239	2-(tert-butyl)-9-butyl-9-hydroimidazo[1,2-a]benzimidazole	269.389	882
1240	methyl_4-(2-{N-[4-(methoxycarbonyl)phenyl]carbamoyl}acetylamino)benzoate	370.361	883
1241	2-(2,5-dimethylpyrrolyl)benzoic_acid	215.251	884
1242	11-phenyl-2,3,4,11,10a-pentahydrobenzo[1',2'-1,2]cyclopenta[3,4-b]quinoline-1,10-dione	327.382	885
1243	[4-(tert-butyl)phenyl]-N-(4-piperidylphenyl)carboxamide	336.476	886
1244	3-[(2,4-dimethylphenyl)amino]-1-phenylpropan-1-one	253.343	886
1245	3-methyl-N-[2-(2-methylindol-3-yl)ethyl]-2-(phenylcarbonylamino)but-2-enamide	375.469	886
1246	3,5-bis[(4-methylphenyl)methylene]-1-methylazaperhydroin-4-one	317.43	887
1247	4-indol-3-yl-6-phenyl-2,4,5-trihydropyridazin-3-one	289.336	888
1248	2-(4-pyridyl)chromeno[4,3-d]pyrimidin-5-one	275.266	890

1249	N-[(1E)-2-(1-methylpyrrol-2-yl)-1-azavinyl]-N'-[(1E)-2-(1-methylpyrrol-2-yl)-1-azavinyl]nonane-1,9-diamide	398.507	890
1250	2-aminophenyl_indolinyl_ketone	238.288	891
1251	ethyl_2-(phenylamino)acetate	179.218	891
1252	2-(1,3-dioxobenzo[c]azolin-2-yl)-N-[2-(2-methylindol-3-yl)ethyl]acetamide	361.399	892
1253	hematoporphyrin	598.697	892
1254	1-{1-[(2-methylphenyl)methyl]pyrazol-3-yl}-1,2,3,4-tetraazole	240.267	894
1255	2,4-dimethylbenzo[h]quinoline	207.274	894
1256	2-{3-[4-(methylethyl)phenyl]-2-(phenylcarbonylamino)propanoylamino}acetic_acid	368.432	895
1257	N-((1E)-2-(2-5,6,7,8-tetrahydronaphthyl)-1-azaprop-1-enyl)-3-pyridylcarboxamide	293.368	895
1258	undecylamine	171.325	895
1259	decyltrimethylazanium_bromide	200.387	897
1260	1,2,3,4,9-pentahydro-4aH-carbazolecarboxylic_acid	215.251	899
1261	2,3,4,9-tetrahydro-4aH-carbazol-1-one	185.225	899
1262	2-{[(3-amino-5-oxo-1-phenyl-1,2-diazolin-4-ylidene)azamethyl]amino}benzoic_acid	323.31	899
1263	(2-methylphenyl){2-[(2-methylphenyl)amino]quinazolin-4-yl}amine	340.427	900
1264	[1-(4-methoxyphenyl)-5-oxopyrrolidin-3-yl]-N-(2,4,6-trimethylphenyl)carboxamide	352.432	900
1265	N-(4-methoxyphenyl){3-[N-(4-methoxyphenyl)-N-methylcarbamoyl]phenyl}-N-methylcarboxamide	404.465	900
1266	3,4-diphenyl-2-(phenylazamethylene)-1,3-thiazoline	328.431	901
1267	(2E)-1-(2,5-dihydroxyphenyl)-3-phenylprop-2-en-1-one	240.258	902
1268	3-(1-ethylindol-3-yl)propanoic_acid	217.267	902
1269	brilliant_blue	749.887	902
1270	(2E)-3-phenyl-1-(2,3,4-trihydroxyphenyl)prop-2-en-1-one	256.257	903
1271	3-[4-(diphenylmethyl)piperazinyl]-1-naphthylazolidine-2,5-dione	475.589	903
1272	N-(2H-benzo[3,4-d]1,3-dioxolen-5-ylmethyl)-N-[1-(4-ethoxyphenyl)-2,5-dioxoazolidin-3-yl]acetamide	410.426	903
1273	ethyl_4-(diethylamino)benzoate	221.299	904
1274	N-[3-((1E)-2-(4-pyridyl)vinyl)phenyl]-2,2-diphenylacetamide	390.484	904
1275	phenylmethyl_4-methyl-2-oxo-6-phenyl-1,3,6-trihydropyrimidine-5-carboxylate	322.363	904
1276	2-(4-methylphenyl)quinoline-4-carboxylic_acid	263.295	905
1277	1,2-dimethylindole	145.204	906
1278	1,5-diaminoanthracene-9,10-dione	238.245	906
1279	3-indol-3-yl-3-hydroisobenzofuran-1-one	249.268	906
1280	benzo[h]quinoline	179.221	906
1281	phenyl[9-benzyl(9-hydro-1,2,4-triazolo[4,5-a]benzimidazol-3-yl)]amine	339.399	906
1282	2-((1E)-2-(3-pyridyl)vinyl)-3-(2-phenylethyl)-3-hydroquinazolin-4-one	353.423	907
1283	3-[(2,5-dimethylphenyl)amino]-5,5-dimethylcyclohex-2-en-1-one	243.348	907
1284	N-{[(dimethylamino)cyclohexyl]methyl}(4-butylcyclohexyl)carboxamide	322.533	907
1285	N-[(4-methylphenyl)phenylmethyl]carboxamide	225.29	908
1286	9-(4-methoxyphenyl)-2,3,4,5,6,7-hexahydroxanthene-1,8-dione	324.376	909
1287	3,4-diacetyl-1-(2-indol-3-ylethyl)-2,5-dimethylpyrrole	322.406	910
1288	3-{4-[2,5-dioxo-1-(2-phenylethyl)azolidin-3-yl]piperazinyl}-1-(2-phenylethyl)azolidine-2,5-dione	488.585	910
1289	methylethyl_3,4-dimethyl-2-oxo-6-phenyl-1,3,6-trihydropyrimidine-5-carboxylate	288.346	910
1290	1-(2-indol-3-ylethyl)-2,5-dimethylpyrrole-3-carboxylic_acid	282.341	911
1291	N-((1Z,3E)-4-phenyl-1-azabuta-1,3-dienyl)-2-(1-methylpyrrol-2-yl)acetamide	267.33	911

1292	4-hydroxy-5,7-dimethyl-3-benzylhydroquinolin-2-one	279.338	912
1293	ethyl_3,5-dimethylpyrrole-2-carboxylate	167.207	912
1294	2-(3,4-dimethoxyphenyl)-7,8-dimethoxychromen-4-one	342.348	913
1295	7-methylindole	131.177	914
1296	phenyl_2,4,5-trimethylpyrrol-3-yl_ketone	213.279	914
1297	2-[(3,4-dimethylphenyl)azamethylene]-3-ethyl-4-phenyl-1,3-thiazoline	308.44	915
1298	4-[(2,4-dimethylphenyl)amino]naphthalene-1,2-dione	277.322	915
1299	5-{[1-(methylpropyl)indol-3-yl]methylene}-1,3-diazolidine-2,4-dione	283.329	915
1300	diethyl_2-{[(2-aminophenyl)amino]methylene}propane-1,3-dioate	278.307	915
1301	N,N-bisbenzyl[4-(2-1,2,3,4-tetrahydroisoquinolylmethyl)phenyl]carboxamide	446.591	915
1302	2-(4-pyrrolidinylphenyl)-1,2,3-trihydroquinazolin-4-one	293.368	919
1303	(2-octylthioethyl)thiocarboxamidine	248.444	920
1304	1,3-dicyclohexylurea	224.345	921
1305	(4-methyl-6-phenylpyrimidin-2-yl)(4,6,8-trimethylquinazolin-2-yl)amine	355.441	922
1306	3-(7,10-dimethoxy-1,2,3,4,10b,4a-hexahydro-6H-benzo[c]chromen-6-yl)phenylamine	339.433	922
1307	ethyl_(2E)-3-[N-(2,2-diphenylacetylamino)carbamoyl]prop-2-enoate	352.389	922
1308	(2E)-1-(4-hydroxyphenyl)-3-phenylprop-2-en-1-one	224.259	923
1309	[2-(2,5-dimethylpyrrolyl)phenyl]methan-1-ol	201.268	923
1310	4-{[4-(diphenylmethyl)piperazinyl]carbonyl}-2-methyl-2-hydrophthalazin-1-one	438.528	923
1311	3,5-bis(1,5-dimethylpyrazol-4-yl)-4-methyl-1-(3-methylphenyl)pyrazole	360.461	924
1312	1-morpholin-4-yl-3,3-diphenylpropan-1-one	295.38	925
1313	2-phenylindole-3-carbaldehyde	221.258	926
1314	N-butyl(4-oxo(3-hydroquinolyl))carboxamide	244.293	926
1315	(3,4-dimethoxyphenyl)-N,N-bis[(3,4-dimethoxyphenyl)methyl]carboxamide	481.544	927
1316	(7-ethyl-4-methylquinazolin-2-yl)(4,6,6-trimethyl(3,6-dihydropyrimidin-2-yl))amine	309.413	927
1317	(2-aminophenyl)-N-(4-methylphenyl)carboxamide	226.277	928
1318	(3-methyl-4-oxo(3-hydrophthalazinyl))-N-(2-phenoxyethyl)carboxamide	323.351	928
1319	N-(3-acetylphenyl)[(4-methylphenyl)amino]carboxamide	268.315	929
1320	2-(3-methylpiperidyl)-4-oxo-5-hydropyridino[1,2-a]pyrimidine-3-carbaldehyde	271.318	930
1321	N-(3-acetylphenyl)(4-methylphenyl)carboxamide	253.3	931
1322	3-aminophenyl_3,4-dimethylphenyl_ketone	225.29	932
1323	8-[(3-methylquinoxalin-2-yl)methylthio]quinoline	317.408	932
1324	N-(3-acetylphenyl)-N'-(3-acetylphenyl)nonane-1,9-diamide	422.523	932
1325	2-phenylquinoline	205.259	934
1326	6-phenyl-1,2,3-trihydroquinolin-4-one	223.274	934
1327	methyl_4-[(3-methyl-2,4-dioxo-1,3-dihydroquinazolinyl)methyl]benzoate	324.335	934
1328	5-methylindole	131.177	935
1329	3-(2,5-dimethylpyrrolyl)benzoic_acid	215.251	936
1330	(2-iminochromen-3-yl)-N-(4-phenoxyphenyl)carboxamide	356.38	939
1331	5-ethyl-1,2-dimethylindole-3-carbaldehyde	201.268	939
1332	3-{N-[(1,3-dioxobenzo[c]azolidin-2-yl)methyl]hexanoylamino}benzoic_acid	394.426	940
1333	5,6-dimethyl-1-[(2,3,5,6-tetramethylphenyl)methyl]benzimidazole	292.423	940
1334	indeno[2,3-b]quinoxalin-11-one	232.241	940
1335	2-(phenylthiomethyl)cyclopent-2-en-1-one	204.286	941
1336	(2-hydroxyphenyl)-N-{4-[(2-hydroxyphenyl)carbonylamino]butyl}carboxamide	328.367	942

1337	3,3,6,6-tetramethyl-9-(4-methylphenyl)-2,3,4,5,6,7,9,10-octahydroacridine-1,8-dione	363.499	942
1338	N-(2,4-dimethylphenyl)(2,4,5-trimethylpyrrol-3-yl)carboxamide	256.347	942
1339	1,3-diphenyl-2-(3-pyridyl)imidazolidine	301.39	943
1340	N-(oxolan-2-ylmethyl)-2-xanthen-9-ylacetamide	323.391	943
1341	(3E)-3-methyl-4-(2-phenylacetylamino)-N-[4-(phenylamino)phenyl]-4-azabut-3-enamide	400.479	944
1342	2,7,8-trimethoxy-12-methyl-12H,5aH-benzo[e]isoindolino[1,2-b]1,3-oxazin-10-one	341.363	944
1343	3-acetyl-2-phenylindole	235.285	944
1344	2-(4-naphthyl-3-1,4-dihydroquinolyl)quinoline	384.479	945
1345	2-methylindolo[2,3-b]quinoxaline	233.272	945
1346	4-(indolinylcarbonyl)phenyl_acetate	281.31	945
1347	ethyl_4-[(1-cyclohexyl-5-oxopyrrolidin-3-yl)carbonylamino]benzoate	358.436	945
1348	2-(1,3-dioxobenzo[c]azolidin-2-yl)-N-{2-[(4-phenylpiperazinyl)carbonyl]phenyl}acetamide	468.511	946
1349	3,5-bis[(2,4-dimethylphenyl)methylene]-1-methylazaperhydroin-4-one	345.483	946
1350	2,5-dimethyl-1-phenylpyrrole-3-carbaldehyde	199.252	948
1351	4-(2,5-dimethylpyrrolyl)benzoic_acid	215.251	948
1352	phenyl-N-[6-(phenylcarbonylamino)(2-pyridyl)]carboxamide	317.346	949
1353	2,4,6,6-tetramethyl-2,4,12b,6a-tetrahydro-6H-pyrimidino[5',4'-6,5]pyrano[3,4-c]chromane-1,3-dione	328.367	950
1354	butyl_4-aminobenzoate	193.245	952
1355	dimethyl_2-[(phenylamino)azamethylene]propane-1,3-dioate	236.227	952
1356	N-((1E)-2-naphthyl-1-azavinyl)(2-aminophenyl)carboxamide	289.336	952
1357	N-[(3-methylphenyl)carbonylamino][2-(phenylcarbonyl)phenyl]carboxamide	358.396	952
1358	3-(4-{[4-(diethylamino)phenyl]methylene}-3-methyl-5-oxo(1,2-diazolinyl))-6-methyl-4H-1,2,4-triazin-5-one	366.422	953
1359	4-((1E)-2-(2H-benzo[3,4-d]1,3-dioxolan-5-yl)-1-azavinyl)-3-methyl-6-phenyl-1,2,4-triazin-5-one	334.334	953
1360	3-hexyloxyphenylamine	193.288	954
1361	N-(2-hydroxyphenyl){3-[N-(2-hydroxyphenyl)carbamoyl]-5-(phenylcarbonylamino)phenyl}carboxamide	467.48	955
1362	N-[(1E)-2-(2-hydroxyphenyl)-1-azavinyl]-3-phenylpropanamide	268.315	955
1363	(5S)-2-methyl-5-(1-methylvinyl)cyclohex-2-en-1-one	150.22	956
1364	2-[(butylamino)methylene]-5-phenylcyclohexane-1,3-dione	271.358	956
1365	5-methyl-2-pyrrolylphenol	173.214	956
1366	7,7-dimethyl-4-naphthyl-3H-4,6,7,8-tetrahydrochromene-2,5-dione	320.387	956
1367	N-(2-methylpropyl)(9-oxofluoren-4-yl)carboxamide	279.338	956
1368	N-(6-methyl(2-pyridyl))(2-(2-pyridyl)(4-quinolyl))carboxamide	340.384	956
1369	N-[2-(3,4-dimethoxyphenyl)-3-phenylpropyl](4-oxo(3-hydrophthalazinyl))carboxamide	443.501	956
1370	2,6-dimethyl-5-(phenylcarbonyl)(3-1,4-dihydropyridyl)_phenyl_ketone	317.387	957
1371	1-acetyl-4-pyrrolylbenzene	185.225	958
1372	2-(3,4-dimethoxyphenyl)-3-methoxychromen-4-one	312.321	958
1373	4,6-dimethyl-2-[(3-methylquinoxalin-2-yl)methylthio]pyrimidine	296.389	958
1374	11-carbamoyl-2-benzylundecanoic_acid	319.443	959
1375	3,3,6,6-tetramethyl-2,3,4,5,6,7-hexahydroacridine-1,8-dione	271.358	961
1376	4-(diethylamino)benzaldehyde	177.246	962
1377	2-[7-methyl-2,4-dioxo-1-phenyl-5-(piperidylcarbonyl)(1,3-dihydropyridino[2,3-d]pyrimidin-3-yl)]-N-phenylacetamide	497.552	964
1378	4-(4-ethylphenyl)-7,7-dimethyl-3H-4,6,7,8-tetrahydrochromene-2,5-dione	298.381	964
1379	N-[3-(dimethylamino)phenyl]-2,2-diphenylacetamide	330.429	964

1380	1,2-dimethylindol-5-yl_acetate	203.24	965
1381	2-(3,4-dimethylphenyl)quinoline-4-carboxylic_acid	277.322	966
1382	2-(4-methylphenyl)quinoline	219.285	966
1383	3-(2-indolinyl-2-oxoethyl)-1,7-dimethyl-5-(piperidylcarbonyl)-1,3-dihydropyridino[2,3-d]pyrimidine-2,4-dione	461.519	966
1384	5-(decanoylamino)pentanoic_acid	271.399	967
1385	1,3-dimethoxy-2-[(3-methylquinoxalin-2-yl)methoxy]benzene	310.352	969
1386	4-methyl-2-(phenylcarbonylamino)benzoic_acid	255.273	970
1387	7,7-dimethyl-4-(2-naphthyl)-3H-4,6,7,8-tetrahydrochromene-2,5-dione	334.414	970
1388	cyclohexyl-N-[2-(2-methylindol-3-yl)ethyl]carboxamide	284.4	971
1389	(2E)-3-(3,4-dimethoxyphenyl)-N-(2-methoxyphenyl)-2-(phenylcarbonylamino)prop-2-enamide	432.475	972
1390	methyl_5,6-diphenyl-1,2,4-triazine-3-carboxylate	291.309	972
1391	2-[(naphthylamino)methylene]cyclohexane-1,3-dione	265.311	973
1392	5-amino-2-(2-naphthyl)benzo[c]azoline-1,3-dione	288.305	974
1393	2-methoxy-1-[6-(4-methoxyphenyl)(3-pyrazolino[5,4-d]1,2,3-triazolin-2-yl)]benzene	323.354	975
1394	ethyl_2-methyl-4-phenylpyrrole-3-carboxylate	229.278	975
1395	N-[2-(2-methylindol-3-yl)ethyl](4-methylphenyl)carboxamide	292.38	976
1396	1,2-dimethyl-1,2,3,4-tetrahydroquinoline-6-carbaldehyde	189.257	977
1397	7,8-dimethoxy-2-phenylchromen-4-one	282.295	977
1398	[(2-phenoxyethyl)amino]-N-(3-{[(2-phenoxyethyl)amino]carbonylamino}phenyl)carboxamide	434.494	978
1399	5-methyl-2-pyrrolylphenylamine	172.229	978
1400	9-ethylcarbazole-3-carboxylic_acid	239.273	978
1401	ethyl_4-(methoxycarbonyl)-1-phenyl-2-pyrazoline-3-carboxylate	276.291	978
1402	N-methyldodecylamine	199.379	980
1403	3-[4-(tert-butyl)phenyl]-3-(phenylcarbonylamino)propanoic_acid	325.407	981
1404	3-cyclopentyl-1-[4-(2-phenoxyacetyl)piperazinyl]propan-1-one	344.453	982
1405	{6-imidazolyl-4-[(4-methylphenyl)amino](1,3,5-triazin-2-yl)}(4-methylphenyl)amine	357.417	983
1406	3-[4-(dimethylamino)phenyl]-2-(4-ethylphenyl)-1,2,3-trihydroquinazolin-4-one	371.481	984
1407	4,5-bis(4-methylphenyl)-2-indol-3-ylimidazole	363.461	984
1408	(15S,2R,14R)-14-acetyl-14-hydroxy-2,15-dimethyl-5-oxotetracyclo[8.7.0.0<2,7>.0<11,15>]heptadeca-6,8-diene	328.45	985
1409	2,3-diphenyl-3-hydroquinazolin-4-one	298.343	985
1410	3-phenyl-1-pyrrolidinyl-2-pyrrolylpropan-1-one	268.358	985
1411	2-[N,N-bisbenzylcarbamoyl]benzoic_acid	345.397	986
1412	2-(N-(2-5,6,7,8-tetrahydronaphthyl)carbamoyl)cyclohexanecarboxylic_acid	301.385	987
1413	N-(2,6-dimethylphenyl)-2-[(2-methylphenyl)amino]acetamide	268.358	987
1414	2-{2,5-dioxo-4-[(2,4,6-trimethoxyphenyl)methylene](1,3-diazolidinyl)}-N-(4-methylphenyl)acetamide	425.44	988
1415	4-(N-hexylcarbamoyl)phenyl_acetate	263.336	988
1416	ethyl_(2E)-3-[N-(2-cyclohex-1-enylethyl)carbamoyl]prop-2-enoate	251.325	990
1417	1,2,2-trimethylpropyl_4-methyl-2-oxo-6-phenyl-1,3,6-trihydropyrimidine-5-carboxylate	316.399	992
1418	2-(2-methoxyphenyl)-N-(3-methylphenyl)acetamide	255.316	992
1419	3-(1-phenylpyridino[3,2-f]quinolin-3-yl)phenol	348.403	992
1420	2-(2-pyridylthiomethylthio)pyridine	234.333	993
1421	2-[(1E)-2-(4-methoxyphenyl)vinyl]benzo[e]1,3-oxazin-4-one	279.295	993
1422	2-{[4-(dimethylamino)phenyl]methylene}-1H-benzo[d]azolin-3-one	264.326	994
1423	3-(4-phenylphenyl)pyrazole	220.273	994
1424	malachite_green	329.464	994

1425	[4-(4,4-diethyl(1H,2H-benzo[d]1,3-oxazin-2-yl))phenyl]diethylamine	338.492	995
1426	7-(diethylamino)chromen-2-one	217.267	995
1427	N-(4-methylphenyl)[2-(4-pyridylcarbonylamino)phenyl]carboxamide	331.373	996
1428	N,N-dimethyldodecylamine	213.406	998
1429	4-phenyl-2-(phenylazamethylene)-3-(2-phenylethyl)-1,3-thiazoline	356.484	999
1430	2-[2-(acetylamino)phenyl]-N-(4-ethylphenyl)-2-oxoacetamide	310.352	1000
1431	1-acetyl-2-methyl-4-(phenylamino)-1,2,3,4-tetrahydroquinoline	280.369	1001
1432	2-methoxy-12-methyl-12H,5aH-benzo[e]isoindolino[1,2-b]1,3-oxazin-10-one	281.31	1001
1433	8-quinolyl_benzoate	249.268	1001
1434	N-[(N-{(1E)-2-[4-(dimethylamino)phenyl]-1-azaprop-1-enyl}carbamoyl)methyl]-2,2-diphenylacetamide	428.533	1001
1435	2-(phenylmethylthio)-1,3-thiazoline	209.324	1002
1436	4-piperidylbenzaldehyde	189.257	1002
1437	3-isoindolin-2-ylphenol	211.263	1003
1438	4-quinoxalin-2-ylphenyl_acetate	264.283	1004
1439	1-benzylindole-3-carbaldehyde	235.285	1006
1440	1-(tert-butyl)-5-(indol-2-ylmethylene)-1,3-dihydropyrimidine-2,4,6-trione	311.34	1007
1441	N-(2,2-dimethylpropyl)(2-hydroxyphenyl)carboxamide	207.272	1008
1442	2-phenylindolizine-1,3-dicarbaldehyde	249.268	1009
1443	ethyl_4-[(3-amino-2-oxochromen-4-yl)amino]benzoate	324.335	1009
1444	4-(phenyldiazenyl)phenol	198.224	1010
1445	methyl_2-amino-5-{[4-amino-3-(methoxycarbonyl)phenyl]methyl}benzoate	314.34	1010
1446	1-methyl-4-(2-oxo-2-(2-5,6,7,8-tetrahydronaphthyl)ethyl)-1,4-dihydroquinoxaline-2,3-dione	348.401	1011
1447	7-hydroquinazolino[4,3-b]quinazolin-8-one	247.256	1011
1448	(3,5-dimethylphenyl)[(4-methoxyphenyl)methyl]amine	241.332	1014
1449	tridecylamine	199.379	1014
1450	4-(5-pentyl-2-pyridyl)benzoic_acid	269.343	1017
1451	5-[(3-methylquinoxalin-2-yl)methoxy]-2H-benzo[d]1,3-dioxolene	294.309	1017
1452	6-methoxy-3-[(2-methylphenyl)carbonyl]chromen-4-one	294.306	1017
1453	N-(3-phenylpropyl)-2-(1,5,7-trimethyl-2,4-dioxo(1,3-dihydropyridino[2,3-d]pyrimidin-3-yl))acetamide	380.446	1017
1454	8-pyrrolylnaphthalen-2-ol	209.247	1018
1455	methyl_4-(N-octylcarbamoyl)butanoate	257.372	1019
1456	2-[2-(2-methylindol-3-yl)ethyl]-4,5,6,7,3a,7a-hexahydroisoindole-1,3-dione	310.395	1020
1457	methyl_2-(2-oxo-3-phenylhydroquinoxalinyl)acetate	294.309	1020
1458	N-((1E)-2-indol-3-yl-1-azavinyl)[4-(tert-butyl)phenyl]carboxamide	319.405	1020
1459	(2E)-1-(2,4-dihydroxyphenyl)-3-phenylprop-2-en-1-one	240.258	1021
1460	N-{(1E)-2-[4-(diethylamino)phenyl]-1-azavinyl}-2-(2-naphthylamino)propanamide	388.511	1021
1461	(4-{[4-(dimethylamino)phenyl][4-(methylamino)phenyl]methylene}cyclohexa-2,5-dienylidene)dimethylamine_chloride	358.505	1022
1462	4-[(3,4-dimethylphenyl)amino]chromen-2-one	265.311	1022
1463	4-methylphenyl_4-(4-{3-[(4-methylphenyl)oxycarbonyl]propanoyl}piperazinyl)-4-oxobutanoate	466.533	1022
1464	4-methoxy-1-[(3-methylquinoxalin-2-yl)methoxy]benzene	280.326	1023
1465	butyl_6-(4-ethylphenyl)-4-methyl-2-oxo-1,3,6-trihydropyrimidine-5-carboxylate	316.399	1024
1466	indeno[3,2-b]quinolin-10-one	231.253	1024
1467	N-(1-acetyl-2-methyl(4-1,2,3,4-tetrahydroquinolyl))(3-methoxyphenyl)-N-benzamide	414.503	1024

1468	(4-aminophenyl)-N-methyl-N-octylcarboxamide	262.394	1025
1469	1,3-dimethyl-5-phenyl-6-benzyl-1,3-dihydropyrrolo[3,4-d]pyrimidine-2,4-dione	345.4	1025
1470	bis(2-ethylhexyl)amine	241.459	1025
1471	N-{[4-hydroxy-2-methyl-5-(methylethyl)phenyl]methyl}(2-hydroxyphenyl)carboxamide	299.369	1026
1472	methyl_blue	755.831	1027
1473	N-(4-{N-[(1E)-2-(3-phenoxyphenyl)-1-azavinyl]carbamoyl}-3-ethoxyphenyl)acetamide	417.463	1027
1474	2-({(1E)-2-[4-(dimethylamino)phenyl]-1-azavinyl}amino)-3-phenyl-3-hydroquinazolin-4-one	383.452	1028
1475	methyl_4-[(N-naphthylcarbamoyl)amino]benzoate	320.347	1028
1476	N-(3,4-dimethoxyphenyl)(3,4,5-triethoxyphenyl)carboxamide	389.447	1032
1477	2-(5,7-dimethyl-2,4-dioxo-1-phenyl(1,3-dihydropyridino[2,3-d]pyrimidin-3-yl))-N-(3-methylbutyl)acetamide	394.472	1033
1478	3-(5,6,7,8,9-pentahydro-4aH-carbazol-9-yl)propanoic_acid	243.305	1035
1479	methyl_3-pyrrolylindole-2-carboxylate	240.261	1035
1480	[4,6-bis(3,5-dimethylpyrazolyl)(1,3,5-triazin-2-yl)]phenylamine	360.421	1036
1481	3-butyl-4-hydroxy-1-phenylhydroquinolin-2-one	293.365	1036
1482	(2E)-3-(3,4-dimethoxyphenyl)-1-phenylprop-2-en-1-one	268.312	1037
1483	1-(2-naphthylmethyl)-2-(naphthylmethyl)benzimidazole	398.506	1037
1484	2-(cyclohexylmethylamino)-4-oxo-5-hydropyridino[1,2-a]pyrimidine-3-carbaldehyde	285.345	1037
1485	5-[(4-ethoxyphenyl)methyl]-1-(4-ethylphenyl)-1,3,5-trihydropyrimidine-2,4,6-trione	366.416	1037
1486	1-(4-methoxyphenyl)-4-{[4-benzylpiperidyl]carbonyl}pyrrolidin-2-one	392.497	1038
1487	(2E)-1,3-bis(2-hydroxyphenyl)prop-2-en-1-one	240.258	1039
1488	isoquinolylthioisoquinoline	288.366	1039
1489	N-[(1E)-2-(2-methylphenyl)-1-azavinyl]-2-hydroxy-2,2-diphenylacetamide	344.412	1039
1490	(2E)-3-[2,5-dimethyl-1-(4-methylphenyl)pyrrol-3-yl]prop-2-enoic_acid	255.316	1041
1491	(3,5-diaminophenyl)-N-(4-methylphenyl)-N-(2-naphthyl)carboxamide	367.449	1041
1492	1-cyclohexyl-2,5-dimethylpyrrole-3-carbaldehyde	205.299	1041
1493	1-methylpyrrolidine	85.1486	428
1494	N-methylpiperidine	99.1754	493
1495	TEMED	116.206	518
1496	N,N-diisopropylethylamine	129.245	546
1497	triallylamine	137.224	575
1498	benzphetamine	239.36	776
1499	cyclobenzaprine	275.393	836
1500	amitriptyline	277.408	851
1501	alverine	281.44	899
1502	pyridine	79.1012	404
1503	2-vinylpyridine	105.139	512
1504	2,4,6-collidine	121.182	544
1505	tetrabutylammonium	242.467	941
1506	androstenedione	286.413	1023
1507	putrescine	88.1522	473
1508	piperidine	85.1486	492
1509	hexmethyleneimine	99.1754	493
1510	3-methyladenine	149.155	493
1511	diallylamine	97.1596	520
1512	diisopropylamine	101.191	524
1513	N-methylbutylamine	87.1644	534
1514	benzylamine	107.155	552

1515	cyclohexylamine	99.1754	554
1516	spermidine	145.247	564
1517	1-phenylethylamine	121.182	578
1518	di-sec-butylamine	129.245	586
1519	spermine	202.342	588
1520	phenethylamine	121.182	595
1521	L-amphetamine	135.208	612
1522	phentermine	149.235	629
1523	mephentermine	163.262	642
1524	dibenzylamine	197.279	690
1525	dicyclohexylamine	181.32	692
1526	octylamine	129.245	727
1527	nonylamine	143.272	785
1528	nortriptyline	263.382	839
1529	maprotiline	277.408	840
1530	apomorphine	267.327	631
1531	levorphanol	257.375	659
1532	levallorphan	283.413	703
1533	pentazocine	271.402	739
1534	4-methylmorpholine	101.148	375
1535	nefopam	253.343	726
1536	dextromethorphan	271.402	761
1537	diphenhydramine	255.359	779
1538	imipramine	280.412	836
1539	1-propyl-4-piperidone	141.213	495
1540	tetramisole	204.289	622
1541	hydrocortisone	362.465	735
1542	cortisone	360.449	745
1543	corticosterone	346.466	836
1544	cortexolone	346.466	856
1545	11a-hydroxyprogesterone	330.466	882
1546	nandrolone	274.402	932
1547	testosterone	288.429	979
1548	deoxycorticosterone	330.466	1001
1549	17a-methyltestosterone	302.456	1039
1550	1,3-dimethyluracil	140.141	487
1551	N-phenylmaleimide	173.171	508
1552	1-methyl-2-piperidone	113.159	534
1553	1,1,3,3-tetramethylurea	116.163	543
1554	N-vinyl-2-pyrrolidinone	111.143	567
1555	N,N-dimethylbenzamide	149.192	646
1556	1-phenyl-2-pyrrolidinone	161.203	715
1557	praziquantel	312.411	936
1558	indole	117.15	843
1559	3-methylindole	131.177	946
1560	carbazole	167.21	1041
1561	1-vinylimidazole	94.1158	474
1562	4-dimethylaminopyridine	122.169	533
1563	m-toluidine	107.155	577
1564	allantoin	158.116	243
1565	2-hydroxypyrimidine	96.0884	282
1566	uracil	112.088	288
1567	dihydrouracil	114.104	290

1568	2,4-diamino-6-hydroxypyrimidine	126.118	348
1569	cytosine	111.103	374
1570	dihydrothymine	128.13	377
1571	thymine	126.115	377
1572	creatinine	113.119	378
1573	2-pyrrolidinone	85.1054	380
1574	2-hydroxypyridine	95.1006	401
1575	3,7-dimethyluric_acid	196.165	464
1576	1,9-dimethyluric_acid	196.165	468
1577	1,3-dimethyluric_acid	196.165	477
1578	1,3,7-trimethyluric_acid	210.192	517
1579	acetanilide	135.165	620
1580	2-hydroxyquinoline	145.16	628
1581	aprobarbital	210.232	696
1582	secbutabarbital	212.248	699
1583	phenytoin	252.272	784
1584	mephobarbital	246.265	811
1585	secobarbital	238.286	821
1586	glutethimide	217.267	825
1587	cyheptamide	237.301	848
1588	cotinine	176.218	505
1589	fentanyl	336.476	777
1590	oxyphenbutazone	324.379	918
1591	3-methyl-2-oxazolidinone	101.105	373
1592	hydrocodone	299.369	617
1593	meperidine	247.336	702
1594	dimefline	323.391	726
1595	mebeverine	429.555	852
1596	drofenine	317.47	929
1597	2-dimethylaminoethanol	89.137	343
1598	triethanolamine	149.189	343
1599	cyclopentanone_oxime	99.1322	555
1600	procyclidine	287.444	825
1601	methcathinone	163.219	601
1602	quinolinic_acid	167.121	252
1603	niacin	123.111	309
1604	3-pyridylacetic_acid	137.138	373
1605	2-dimethylaminobenzoic_acid	165.191	498
1606	3-dimethylaminobenzoic_acid	165.191	618
1607	4-dimethylaminobenzoic_acid	165.191	731
1608	dimethylglycine	103.121	246
1609	betaine	118.155	246
1610	2-aminopyrimidine	95.1036	445
1611	proflavine	209.25	645
1612	aminoquinoline	144.176	691
1613	ethidium	314.409	769
1614	4-aminophenol	109.127	418
1615	p-anisidine	123.154	551
1616	xylazine	220.332	685
1617	dopamine	153.18	506
1618	dobutamine	301.385	638
1619	mescaline	211.26	595
1620	methoxyphenamine	179.261	651

1621	desipramine	266.385	827
1622	promazine	284.418	812
1623	thiethylperazine	399.611	916
1624	tripelennamine	255.362	722
1625	detomidine	186.256	694
1626	oxindole	133.149	607
1627	doxylamine	270.374	646
1628	R-apocodeine	281.354	694
1629	choline	104.172	343
1630	2-amino-3-phosphonopropionic_acid	169.074	222
1631	taurine	125.142	236
1632	L-aspartic_acid	133.104	241
1633	diaminopimelic_acid	190.199	250
1634	D-glutamic_acid	147.13	253
1635	L-proline	115.132	260
1636	L-alpha-aminobutyric_acid	103.121	278
1637	guanidinosuccinic_acid	175.144	283
1638	2-aminoisobutyric_acid	103.121	287
1639	aminoadipic_acid	161.157	288
1640	guanidoacetic_acid	117.107	291
1641	creatine	131.134	297
1642	pipecolinic_acid	129.158	310
1643	D-valine	117.147	326
1644	L-lysine	146.189	326
1645	L-arginine	174.202	340
1646	homo-L-arginine	188.229	394
1647	5-aminopentanoic_acid	117.147	449
1648	2-phenylglycine	151.165	458
1649	L-isoleucine	131.174	482
1650	4-guanidinobutanoic_acid	145.161	488
1651	leucine	131.174	494
1652	L-norleucine	131.174	502
1653	6-aminocaproic_acid	131.174	506
1654	N-methyl-a-aminoisobutyric_acid	117.147	508
1655	beta-leucine	131.174	521
1656	phenylalanine	165.191	521
1657	2-(N-morpholino)ethanesulfonic_acid	195.233	243
1658	morpholinopropane_sulfonic_acid	209.26	246
1659	3-amino-1,2-propanediol	91.1096	321
1660	D-glucosamine	179.172	322
1661	trizma	121.136	326
1662	2-amino-2-methyl-1,3-propanediol	105.136	347
1663	heptaminol	145.244	555
1664	N-benzylethanolamine	151.208	560
1665	ephedrin	165.235	599
1666	scopolamine	303.357	605
1667	oxycodone	315.368	607
1668	ethylmorphine	321.459	625
1669	atropine	289.374	636
1670	tramadol	263.379	669
1671	cinchonine	294.396	633
1672	hydroxybutorphanol	343.465	625
1673	lidocaine	234.341	656

1674	ropivacaine	274.405	706
1675	bupivicaine	288.432	750
1676	2-oxazolidinone	87.078	277
1677	urethane	89.0938	400
1678	meprobamate	218.252	662
1679	acetaminophen	151.165	494
1680	hydroxyphenytoin	268.271	668
1681	niacinamide	122.126	432
1682	1-phenyl-3-pyrazolidinone	162.191	642
1683	1-methylnicotinamide	137.161	366
1684	methyprylon	183.25	649
1685	lipoamide	205.333	713
1686	3,5-diaminobenzoic_acid	152.152	388
1687	m-aminobenzoic_acid	137.138	503
1688	kynurenic_acid	189.17	509
1689	N-phenylglycine	151.165	599
1690	N-2-carboxyphenylglycine	195.174	656
1691	3,5-dimethylanthranilic_acid	165.191	745
1692	N-ethylanthranilic_acid	165.191	776
1693	2-ethylanilinoacetic_acid	179.218	794
1694	nalidixic_acid	232.238	790
1695	diphenylamine-4-sulfonic_acid	249.284	531
1696	indole-3-carboxylic_acid	161.16	640
1697	indole-3-acetic_acid	175.187	678
1698	3-indolepropionic_acid	189.213	741
1699	1-methyl-4-imidazoleacetic_acid	140.141	348
1700	tolmetin	257.288	871
1701	L-homoserine_lactone	101.105	338
1702	glycine_methyl_ester	89.0938	361
1703	glycine_ethyl_ester	103.121	475
1704	L-phenylalanine_ethyl_ester	193.245	641
1705	methylphenidate	233.31	680
1706	3-methoxytyramine	167.207	547
1707	doxapram	378.513	695
1708	alfentanil	416.522	754
1709	ethamivan	223.271	673
1710	methyl-2-oxo-1-pyrrolidine-acetate	157.169	508
1711	pyridoxal	167.164	484
1712	pyridoxine	169.18	510
1713	acepromazine	326.456	795
1714	propiomazine	340.482	848
1715	phosphorylcholine	184.152	234
1716	prostaglandin_B1	336.47	978
1717	resorufin	213.192	680
1718	benzocaine	165.191	767
1719	1-methylhistamine	125.173	527
1720	tryptamine	160.218	613
1721	trimethoprim	290.321	624
1722	pyrilamine	285.388	719
1723	7-methyladenine	149.155	475
1724	6-dimethylaminopurine	163.182	536
1725	ipratropium	332.462	659
1726	indoxyl_acetate	175.187	834

1727	benzydamine	309.41	807
1728	caffeine	194.193	548
1729	2,4-quinoline_diol	161.16	609
1730	L-serine_methyl_ester	119.12	375
1731	nadolol	309.405	599
1732	metoprolol	267.367	660
1733	toliprolol	223.314	692
1734	oxprenolol	265.352	711
1735	bisoprolol	325.447	717
1736	propranolol	259.347	742
1737	betaxolol	307.432	763
1738	bitolterol	461.557	993
1739	epinephrine	183.207	470
1740	phenylephrine	167.207	531
1741	albuterol	239.314	554
1742	ritodrine	287.358	611
1743	norbutorphanol	259.347	616
1744	ractopamine	301.385	638
1745	L-cystine	240.292	243
1746	allocystathionine	222.259	246
1747	homocystine	268.345	324
1748	L-methionine	149.207	339
1749	L-dopa	197.19	342
1750	L-tyrosine	181.191	465
1751	o-tyrosine	181.191	513
1752	phosphoserine	185.073	207
1753	5-aminolevulinic_acid	131.131	347
1754	N-acetyl-L-glutamine	188.183	277
1755	L-pyroglutamic_acid	129.115	287
1756	isovalerylglycine	159.185	523
1757	suberylglycine	231.248	529
1758	phenylacetyl_L-glutamine	264.28	550
1759	phenylacetylglycine	193.202	565
1760	nalorphine	311.38	589
1761	naloxone	327.379	591
1762	naltrexone	341.406	606
1763	nalbuphine	357.449	623
1764	quinidine	324.422	652
1765	N-acetylputrescine	130.189	451
1766	MEGX	206.287	624
1767	prilocaine	220.314	655
1768	3-hydroxylidocaine	250.34	586
1769	fenspiride	260.335	621
1770	7-methylxanthine	166.139	444
1771	theobromine	180.166	490
1772	paraxanthine	180.166	514
1773	etamiphylline	279.341	564
1774	propentofylline	306.364	733
1775	hepes	238.301	266
1776	fexofenadine	501.664	828
1777	glycerophosphocholine	258.231	235
1778	L-carnitine	162.208	344
1779	procaine	236.313	605

1780	anileridine	352.475	726
1781	tetracaine	264.367	773
1782	methylhippuric_acid	193.202	617
1783	colchicine	399.443	673
1784	remifentanil	376.452	699
1785	imidazoleacetic_acid	126.115	339
1786	urocanic_acid	138.126	481
1787	5-hydroxyindoleacetic_acid	191.186	542
1788	xanthurenic_acid	205.17	484
1789	1-aminoanthraquinone-2-sulfonic_acid	303.289	610
1790	2-methoxyanilinoacetic_acid	181.191	670
1791	adenine	135.128	465
1792	6-methyladenine	149.155	508
1793	benzoylecgonine	289.33	610
1794	L-acetylcarnitine	204.245	486
1795	stanozolol	328.497	966
1796	5-hydroxytryptophol	177.202	530
1797	histamine	111.146	524
1798	reserpine	608.687	891
1799	N,N-diallyltartardiamide	228.247	461
1800	buspirone	385.508	734
1801	serine	105.093	238
1802	hydroxyproline	131.131	241
1803	tricine	179.172	242
1804	L-threonine	119.12	245
1805	L-homoserine	119.12	249
1806	morphine-3-glucuronide	461.468	499
1807	8-hydroxy-2-deoxyguanosine	283.243	463
1808	cytidine	243.219	468
1809	deoxycytidine	227.219	485
1810	thymidine	242.231	634
1811	riboflavin	376.368	537
1812	biotin	244.308	556
1813	penicillin_G	334.389	744
1814	nicotinuric_acid	180.163	443
1815	N-carbobenzyloxy-L-aspartic_acid	267.238	639
1816	3-methylhistidine	169.183	318
1817	1-methylhistidine	169.183	335
1818	tryptophan	204.228	554
1819	leucylproline	228.291	570
1820	lisinopril	405.493	580
1821	xanthine	152.112	326
1822	melatonin	232.282	662
1823	pyridoxamine	168.195	501
1824	isoxsuprine	301.385	707
1825	L-thyronine	273.288	589
1826	benzyl_L-glutamate	237.255	597
1827	omeprazole	345.415	676
1828	yohimbine	354.448	684
1829	dibucaine	343.468	846
1830	procainamide	235.328	561
1831	indolelactic_acid	205.213	622
1832	L-asparagine	132.119	240

1833	glutamine	146.146	246
1834	citrulline	175.187	253
1835	glycyl-glycine	132.119	293
1836	N6-acetyl-L-lysine	188.226	323
1837	N-alpha-acetyllysine	188.226	383
1838	glycyl-L-leucine	188.226	565
1839	prolylisoleucine	228.291	574
1840	L-aspartyl-L-phenylalanine	280.28	577
1841	diprotin_A	341.45	617
1842	pantothenic_acid	219.237	473
1843	taurocholic_acid	515.704	657
1844	glycocholic_acid	465.629	789
1845	glycodeoxycholate	449.629	937
1846	histidine	155.156	325
1847	L-5-hydroxytryptophan	220.227	501
1848	5-methoxytryptophan	234.254	554
1849	bambuterol	367.444	685
1850	deacetyldiltiazem	372.481	741
1851	7-methylguanine	165.154	480
1852	1-methylguanine	165.154	481
1853	adenosine	267.244	489
1854	pindolol	248.324	629
1855	hypoxanthine	136.113	320
1856	4-aminohippuric_acid	194.19	463
1857	formoterol	344.41	668
1858	cytidine_monophosphate	323.199	248
1859	dCMP	307.199	285
1860	flavin_mononucleotide	456.348	479
1861	procaterol	290.361	568
1862	labetalol	328.41	714
1863	carteolol	292.377	601
1864	guanine	151.127	311
1865	fenbendazole	299.347	875
1866	biocytin	372.482	522
1867	5-methylthioadenosine	297.331	550
1868	ADP	427.204	213
1869	adenosinemonophosphate	347.224	283
1870	adenylsuccinic_acid	463.297	300
1871	dAMP	331.224	301
1872	cyclic_AMP	329.209	416
1873	citicoline	489.336	244
1874	nicotinamideribotide	335.23	256
1875	amoxicillin	365.403	547
1876	guanosine	283.243	449
1877	deoxyguanosine	267.244	458
1878	carnosine	226.235	411
1879	7-methylguanosine	298.278	496
1880	inosinic_acid	348.209	234
1881	GMP	363.223	246
1882	2-deoxyguanosine-5-monophosphate	347.224	284

Appendix 2. The scripts used for filtering

2.1 log D similarity searching

The absolute difference of the log D molecular descriptor values of compounds in database to the target compound was calculated as a filter. The threshold used for the log D approach was 0.2, compounds with the absolute difference of the log D less than 0.2 were used as the training set to build QSRR models for the target compound [2]. The scripts are supposed to run on a Matlab platform and are listed below.

```
clear all
clc
values=xlsread('lgD_88 compounds.xlsx','lgD','a1:a88'); % put the right name of the Excel name,
the worksheet name and the matrix contains the values of the log D for all the compounds.
number=zeros(max(size(values)),1);
selected_ID=zeros(max(size(values)),max(size(values)));
selected_values=zeros(max(size(values)),max(size(values)));
difference=zeros(max(size(values)),max(size(values)));
maxdiff=0.5; % put the threshold of the log D filter here
ID=1:max(size(values));
ID=ID.';
for ref=1:max(size(values))
j=0;
      for i=1:max(size(values))
         difference(i,ref)=abs(values(ref)-values(i));
         if abs(difference(i,ref))<maxdiff && abs(difference(i,ref))>0
            j=j+1;
            number(ref,1)=j;
         end
      end
   [sort_difference(:,ref),index]=sort(difference(:,ref));
   sort_ID(:,ref)=ID(index);
   sort_values(:,ref)=values(index);
end
for ref=1:max(size(values))
   number(ref,1)=0;
   j=2;
   while sort_difference(j,ref)<maxdiff || j>max(size(values))
      number(ref,1)=j-1;
      j=j+1;
   end
   selected_ID(1:number(ref,1)+1,ref)=sort_ID(1:number(ref,1)+1,ref);
end
```

2.2 log P similarity searching

The ratio of the log P molecular descriptor values of compounds in database to the target compound was also evaluated as a filter. The threshold used for the log P approach was 1.2, compounds with the ratio of the log P less than 1.2 were used as the training set to build QSRR models for the target compound. The scripts are supposed to run on a Matlab platform and are listed below.

```
Y=xlsread('lgP_88 compounds.xlsx','lgP','a1:a88'); % put the right name of the Excel name, the
worksheet name and the matrix contains the values of the log P for all the compounds.
ID=1:max(size(Y));
ID=ID.';
ratio=zeros(max(size(Y)),max(size(Y)));
selected_ID=zeros(max(size(Y)),max(size(Y)));
mink = 0;
maxratio = 1.1; % put the threshold of the log P filter here
%ranking k-values:
for unseen=1:max(size(Y))
  if Y(unseen)>mink
    for i=1:max(size(Y))
      if Y(i)<mink
        ratio(i,unseen)=999;
      else
        ratio(i,unseen)=Y(i)./Y(unseen);
      end
      %ratio must be positive
      if ratio(i,unseen)<1
        ratio(i,unseen)=Y(unseen)./Y(i);
      end
    end
  else
    ratio(:,unseen)=333;
  end
  [sort_ratio(:,unseen),index]=sort(ratio(:,unseen));
  sort_ID(:,unseen)=ID(index);
end
```

2.3 Local training set for single target compound

This local QSRR model provides a separate prediction of retention for each target compound. Using the scripts, predictions for all the target compounds can be performed on the Matlab platform automatically. To perform local QSRR modelling, scripts are listed below.

```
clear all
clc
ID=xlsread('D:\yabin\5 new\5 new compounds.xlsx','ID','a1:ew11'); % put the id of compounds
in training set of target compounds.
X=xlsread('D:\yabin\5 new\5 new compounds.xlsx','descriptors','a2:dv154'); % put the matrix of
the molecular descriptors of all the compounds
Y=xlsread('D:\yabin\5 new\5 new compounds.xlsx','eta','a2:a154'); % put the responses (solute
coefficients or retention times) of all the compounds.
[niets, descriptorsAll] =xlsread('D:\yabin\5 new\5 new compounds.xlsx','descriptors','a1:dv1'); %
put the matrix of the names of the selected molecular descriptors.
meanAbsError=zeros(size(X,1),1);
meanError=zeros(size(X,1),1);
ErrorPLS=zeros(size(X,1),1);
Q2_max_PLS=zeros(size(X,1),1);
Error=zeros(size(X,1),5);
Error_all=zeros(size(X,1),6);
medianError=zeros(size(X,1),1);
origid=zeros(size(X,2),size(X,1));
for unseen=1:size(X,1) % put the initial number of the compound need to be calculated.
```

```
      Ind=ID(:,unseen);
       sz=sum(Ind~=0);
       if sz > 4 % min training set size ≥ 5
       XX=X(Ind(1:sz),:);
       yy=Y(Ind(1:sz));
       id_ct= find( arrayfun( @(c) numel( unique( XX(:, c) )), 1:size(XX, 2 )) > 16 );
       descriptors=descriptorsAll(id_ct);
       descriptors=descriptors.';
       Xcal=XX(2:end,id_ct);
       ycal=yy(2:end);
       Xunseen=XX(1,id_ct);
       yunseen=yy(1);
       [PLS_LOO_ref,n]=plscv_Eva(Xcal,ycal,5,size(Xcal,1),'autoscaling',1,2);
       PLS_ref_nlv=pls(Xcal,ycal,n,'autoscaling'); % n = opt. # LV obtained
       PLS_pred_nlv=plstest(PLS_ref_nlv,Xunseen,yunseen,n); % it gives prediction error for
target.
       ErrorPLS(unseen,1) = [PLS_pred_nlv.error].';
       Q2_max_PLS(unseen,1) = [PLS_LOO_ref.Q2_max].';
       gapls_loop
      var=logical(variable_index);
       loop_CV;
       F1 = struct('Q2', {F(1:5).Q2}, 'optLV', {F(1:5).optLV});
       pred_loop
       Error(unseen,1:5)= [P.error]; % prediction error for 5-times GA-PLS
       meanError(unseen,1)=mean([P.error].'); % mean error
       meanAbsError(unseen,1)=mean(abs([P.error].')); % mean absolute error
       medianError(unseen,1)=median([P.error].'); % median error
       freq=sum(variable_index == 1,2);
       id=find(freq>0);
       OID=id_ct(id);
       selected_desc=[num2cell(OID).' descriptorsAll(OID).'];
       origid(:,unseen)=ismember((1:size(X,2)),OID).';
       filename = [ '5 new compounds Tanimoto_' num2str(unseen) '.mat' ];
       save(filename);
catch
     filename = [ '5 new compounds Tanimoto_' num2str(unseen) '.mat' ];
       save(filename);
 end
   else
       ErrorPLS(unseen,1) = 999;
       Q2_max_PLS(unseen,1) = 999;
       Error (unseen,1:5)= 999.*ones(1,5);
       meanAbsError(unseen,1)= 999;
       medianError(unseen,1)= 999;
       meanError(unseen,1)= 999;
       id(unseen,1)=999;
       Error_all(unseen,:)=[Q2_max_PLS(unseen) ErrorPLS(unseen) abs(ErrorPLS(unseen))
meanAbsError(unseen) medianError(unseen) meanError(unseen) ];
       end
       Error_all(unseen,:)=[Q2_max_PLS(unseen) ErrorPLS(unseen) abs(ErrorPLS(unseen))
meanAbsError(unseen) medianError(unseen) meanError(unseen) ];
end
```

2.4 Local training set for a group of target compounds

These QSRR models were formed for the prediction of solute coefficients and retention times [3]. Unlike above-mentioned local QSRR models, each target compound has its own specific training set to build its own QSRR model. Here, predictions were performed for a group of compounds that generated using different filtering approaches. To perform QSRR modelling manually, scripts are listed below.

```
[Train,Test,Text,Qual,Xtr,Xts,Xtx] = main_mt(name,number); % split compounds into training and test set, 70% in training set and 30% in test set. Put the name of the worksheet which contains the variables (retention times) of compounds, put the number of compounds in training set as well.
```

Training sets are used to build QSRR models, test sets are taken to verify the predictive ability of the formed QSRR models. To run the codes on the Matlab, five matrices need to be prepared first. The molecular descriptors of compounds in training set, the retention times (or solute coefficients) of compounds in training set, the molecular descriptors of compounds in test set and the retention times (or solute coefficients) of compounds in test set. The last matrix needed is the one which contains the parameters of the genetic algorithm (GA). The first four matrices are named Xcal, ycal, Xtest and ytest, respectively.

```
Xcal=xlsread('   ','   ','   ');
ycal=xlsread('   ','   ','   ');
Xtest=xlsread('   ','   ','   ');
ytest= xlsread('   ','   ','   '); % prepare the file contains the above-mentioned data, put the corresponding names, values, matrix of locations correctly in spaces.
options = ga_options('pls','auto',100); % parameters of the GA are embedded.
options.max_param = 5; % y-randomization to determine optimum # evaluations.
ga=ga_model(Xcal,ycal,options,n); % enter the optimum # obtained into options.
ga1 = ga_model(Xcal,ycal,options,0) % build GA-PLS models 5 times.
ga2 = ga_model(Xcal,ycal,options,0)
ga3 = ga_model(Xcal,ycal,options,0)
ga4 = ga_model(Xcal,ycal,options,0)
ga5 = ga_model(Xcal,ycal,options,0)
ga1_step_selection = sortrows(ga1.step_selection',-2) % 5 times step selection.
ga2_step_selection = sortrows(ga2.step_selection',-2)
ga3_step_selection = sortrows(ga3.step_selection',-2)
ga4_step_selection = sortrows(ga4.step_selection',-2)
ga5_step_selection = sortrows(ga5.step_selection',-2)
ps=ga_postselection_short(Xcal,ycal,m,n,nan) % choose the subset and run post selection.
ps_step_select=sortrows([(1:size(ps.as.step_selection,1))', ps.as.step_selection],-3); % select final subset
var_sel_nlv=ps.as.var_selected{1,n}
desc_sel_name=descriptors(var_sel_nlv) % extract names of the selected descriptors.
PLS_nLV=plsfit(Xcal(:,var_sel_nlv),ycal,n,'auto') % using optimum # of LV to build PLS model for training set.
pred_reg_param = calc_external_reg_param(ycal,ytest,pred_nLV.yc) % using the same optimum # of LV to build PLS model for test set.
```

Appendix 3. User manual for QSRR modelling

In the present study, a variety of software tools were employed in the QSRR modelling, for processes such as conformer searching, molecular descriptor calculation, GA-PLS on the Matlab platform and other steps. Here, a step-by-step guide is presented to show the exact procedures that have been used. All scripts used (bash scripts and Matlab scripts etc.) are included in Appendix 2.

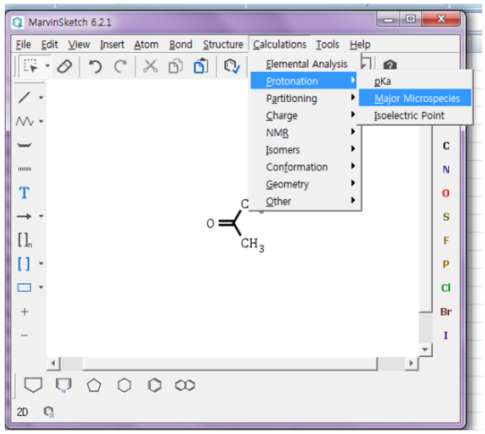
3.1 Dragon for the generation of molecular descriptors

Obtaining the structures of compounds is the first step. Marvin Sketch (ChemAxon) was employed in the present study to draw the structures of compounds directly. Structures of compounds need to be saved as .sdf files.

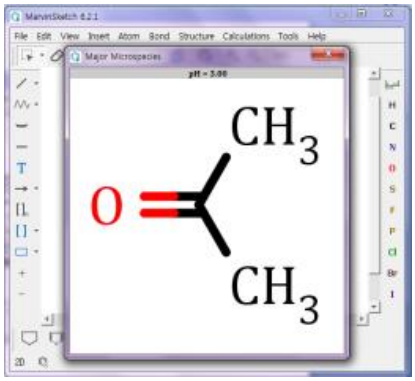
To find the protonation state of each molecule, first decide a pH value that is relevant to the situation, then create folders for each charge state: minus one, neutral, etc. After that, open Marvin Sketch and follow the next instructions:

1. Open the .sdf file.

2. Click on “Calculations”, then “Protonation”, then click on “Major Microspecies”



3. Choose the pH value, then the protonation state of the compound at that pH will be displayed.



Another method of finding the state is to click on “Calculations”, then “Protonation”, then click on “pKa” and choose “Enter”. The program will then show a graph of the major protonation states available for the compound. Once the major microspecies are found under a user-defined pH, correct the structure and add charges if needed using the charge buttons (+ and –) in Marvin Sketch. Finally, save the structure of each compound to the folder with the correct charge state.

Conformer searching was performed using MMFF94 to find the 50 lowest energy conformers for each compound. Conformers were calculated using a forcefield for the appropriate solvent, but this also could be done in the gas phase if required. For the searching of conformers, the program Balloon and a bash script written by Georg Schuster was utilised. The scripts needed for conformer searching in acetonitrile are called:

"config_search_solvent.bat "

"config_search_solvent.sh"

To perform the calculation, put a copy of the scripts into the charge state folders. Double click the .bat script and follow the instructions given in the command line. The commands are shown below:

–1

OPT

2

50

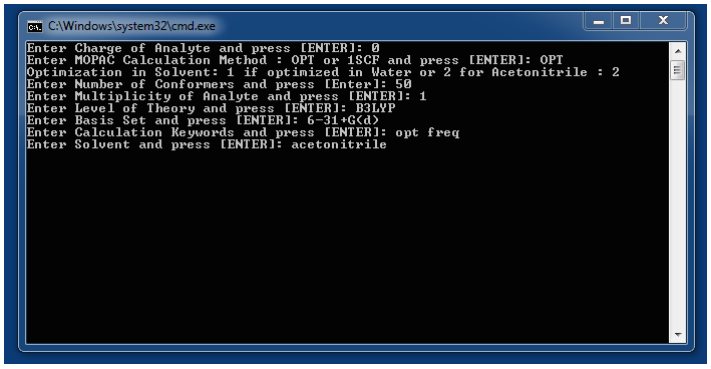
1

B3LYP

6-31+G(d)

opt freq

acetonitrile



Output from conformer searching: each compound will be assigned into a separate folder, and the 50 different conformers will be kept in that folder. Also, an excel spreadsheet listing the conformers in order of energy with the lowest energy (most stable) conformer placed first on the list will be generated. It is worth pointing out that a new folder for the most stable conformer (the one to use for descriptor calculations) will be given within that molecule folder. In each molecule folder, there are a few different files including:

.arc

.aux

.com – this file is for use in DFT calculations using gaussian

.dat

.mol2 – this file is to use for descriptor calculations (PM7 geometry) using Dragon

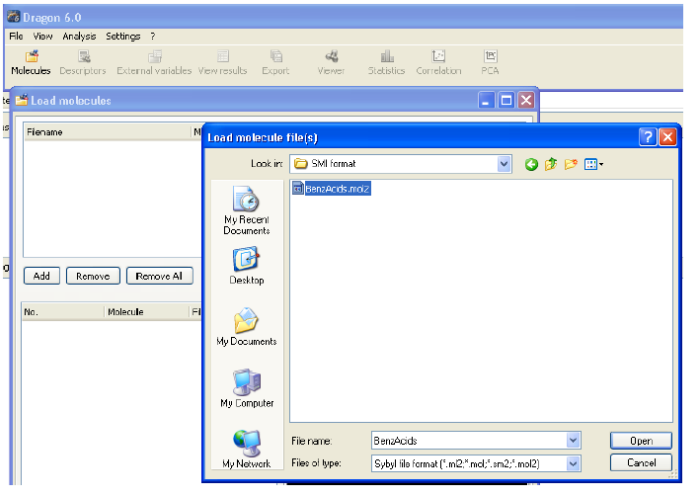
.out – this is the output file from the geometry optimisation in MOPAC

.xyz – this file can be read by Avogadro to check the geometry

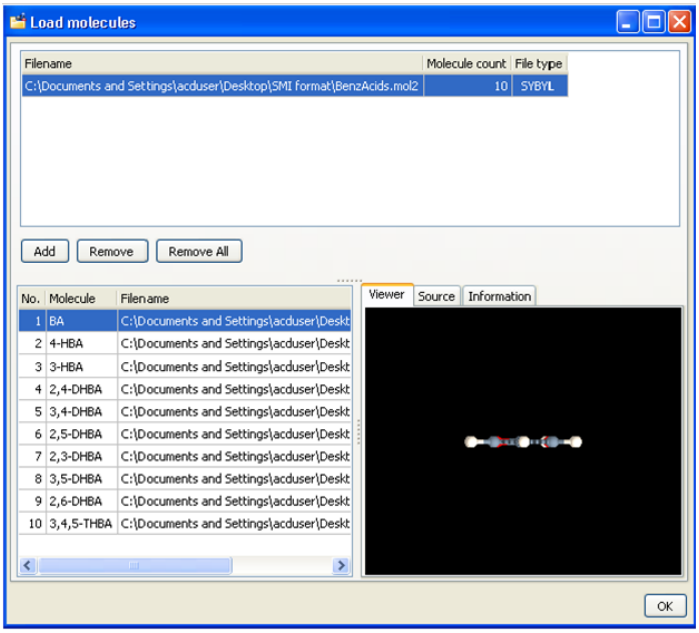
_1scf.mol2 – this is the one you use if the geometry of the optimised structure is wrong.

Once the step of conformer searching is done, the molecular descriptors then can be generated using Dragon 6:

1. Launch Dragon 6 and select "Molecules" and click on "Add" to import the files. Locate the ".mol2" files obtained from the conformer search (PM7 method).

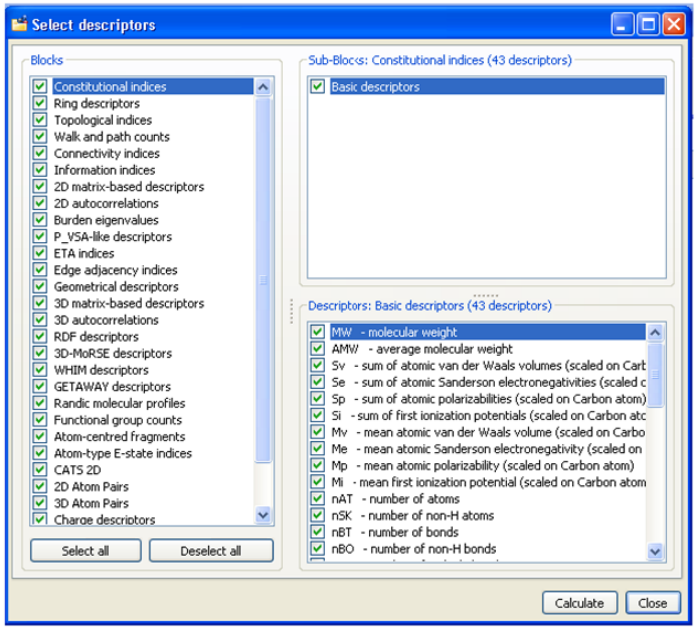


2. The following window will appear, in which you can see your data file. Click OK to upload your dataset into Dragon.

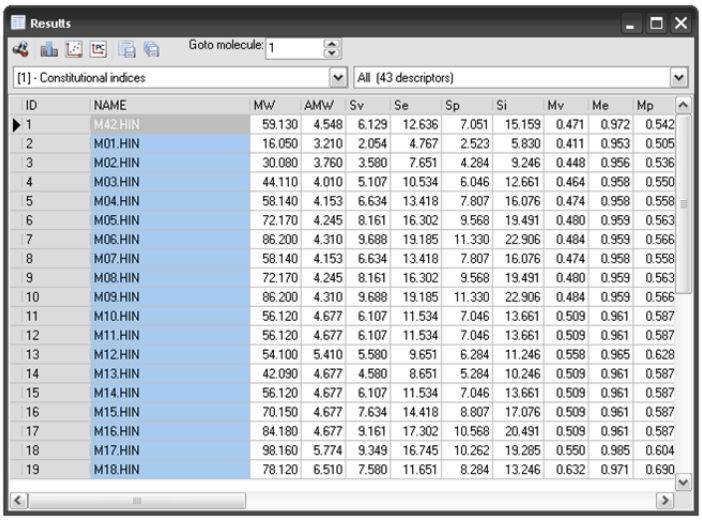


3. When the "Descriptors" tab in the toolbar is activated, click on it to start.

4. A variety of molecular descriptors (shown as blocks) will be displayed, click on "Select All" to choose all the available descriptors and then click "Calculate".

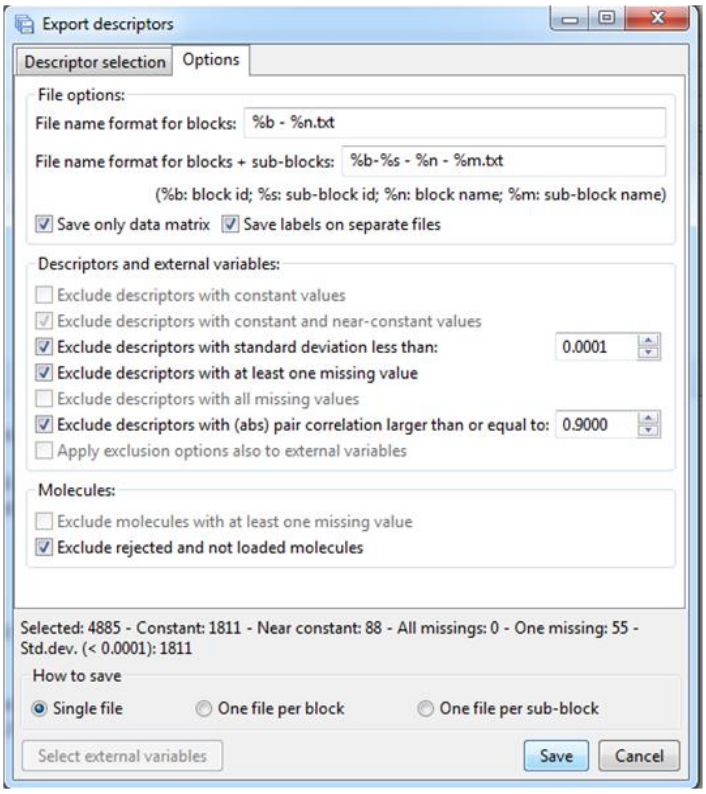


5. Once the calculation finished, click on "View Results" tab to view the generated molecular descriptors. The first column is the number of compounds, in each row, the values of molecular descriptors are listed for each compound.



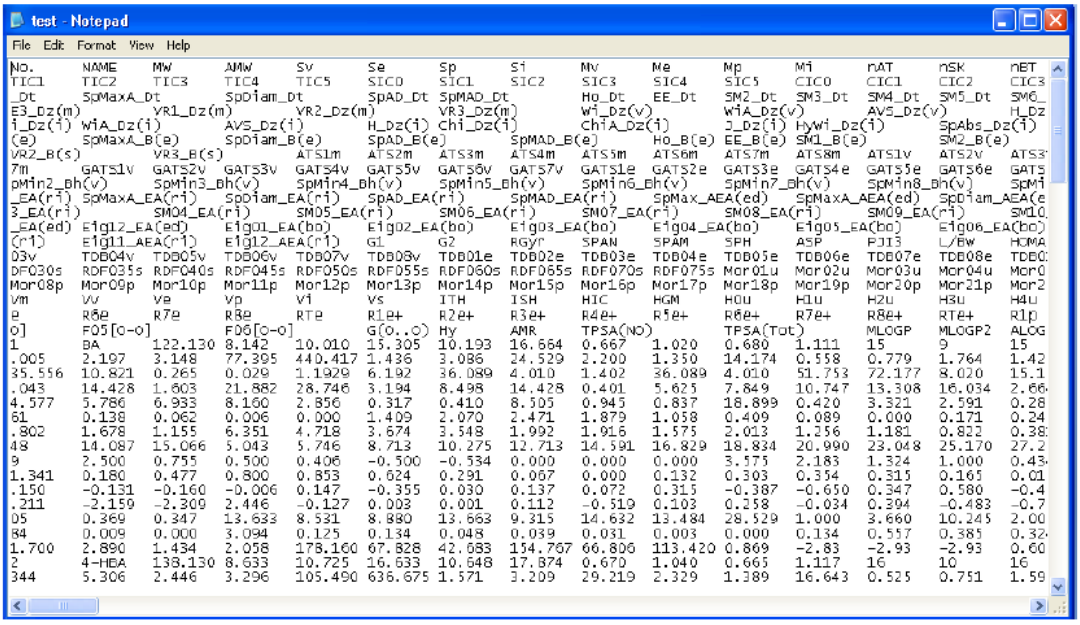
ID	NAME	MW	AMW	Sv	Se	Sp	Si	Mv	Me	Mp
1	M42.HIN	59.130	4.548	6.129	12.636	7.051	15.159	0.471	0.972	0.542
2	M01.HIN	16.050	3.210	2.054	4.767	2.523	5.830	0.411	0.953	0.505
3	M02.HIN	30.080	3.760	3.580	7.651	4.284	9.246	0.448	0.956	0.536
4	M03.HIN	44.110	4.010	5.107	10.534	6.046	12.661	0.464	0.958	0.550
5	M04.HIN	58.140	4.153	6.634	13.418	7.807	16.076	0.474	0.958	0.558
6	M05.HIN	72.170	4.245	8.161	16.302	9.568	19.491	0.480	0.959	0.563
7	M06.HIN	86.200	4.310	9.688	19.185	11.330	22.906	0.484	0.959	0.566
8	M07.HIN	58.140	4.153	6.634	13.418	7.807	16.076	0.474	0.958	0.558
9	M08.HIN	72.170	4.245	8.161	16.302	9.568	19.491	0.480	0.959	0.563
10	M09.HIN	86.200	4.310	9.688	19.185	11.330	22.906	0.484	0.959	0.566
11	M10.HIN	56.120	4.677	6.107	11.534	7.046	13.661	0.509	0.961	0.587
12	M11.HIN	56.120	4.677	6.107	11.534	7.046	13.661	0.509	0.961	0.587
13	M12.HIN	54.100	5.410	5.580	9.651	6.284	11.246	0.558	0.965	0.628
14	M13.HIN	42.090	4.677	4.580	8.651	5.284	10.246	0.509	0.961	0.587
15	M14.HIN	56.120	4.677	6.107	11.534	7.046	13.661	0.509	0.961	0.587
16	M15.HIN	70.150	4.677	7.634	14.418	8.807	17.076	0.509	0.961	0.587
17	M16.HIN	84.180	4.677	9.161	17.302	10.568	20.491	0.509	0.961	0.587
18	M17.HIN	98.160	5.774	9.349	16.745	10.262	19.285	0.550	0.985	0.604
19	M18.HIN	78.120	6.510	7.580	11.651	8.284	13.246	0.632	0.971	0.690

6. To save the output, click on "Export" and then click on "Option" tab and select the options as seen below then click on "Save" to save the results as a text file. To avoid chance correlation, some molecular descriptors need to be excluded, such as constant or near constant values, descriptors with a standard deviation less than 0.0001, descriptors strongly correlated to others (coefficient of determination > 0.90), and descriptors not available for all compounds.

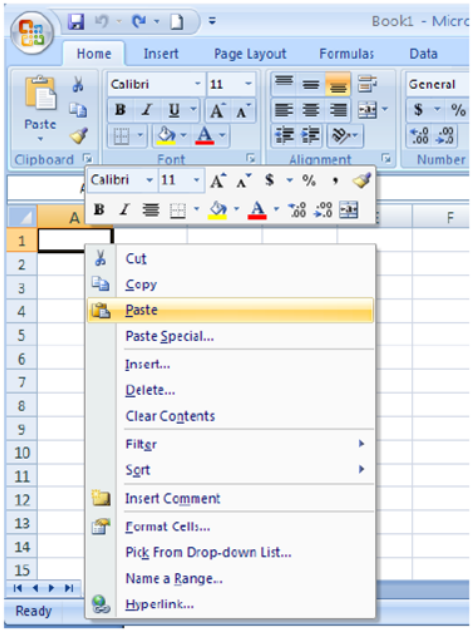


7. At this point the molecular descriptors are ready to export. The following message box will appear which shows the number of descriptors that have been exported.

8. The last step is transfer the generated molecular descriptors from the text file to a MS Excel file for its subsequent use Open the saved text file.



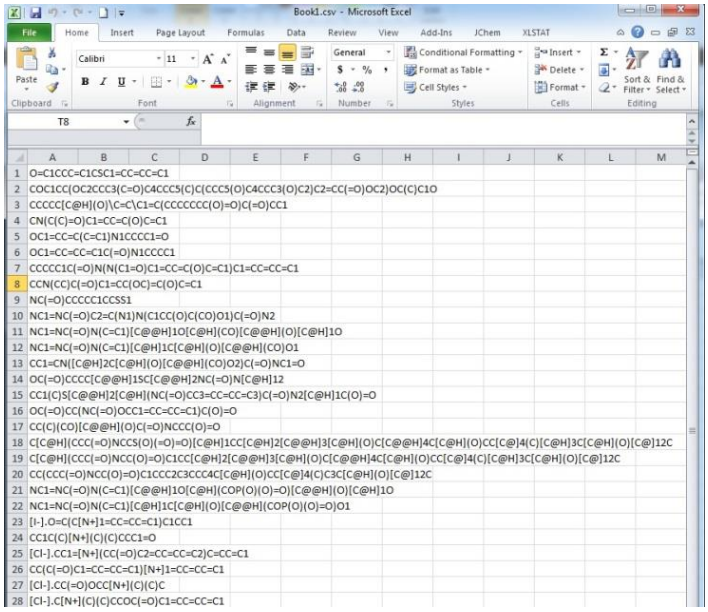
9. Transfer all the data using the functions "copy" and “paste”, and save the Excel file.



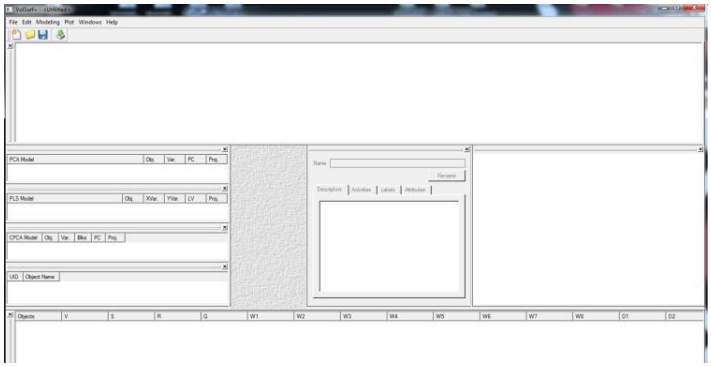
3.2 VolSurf+ for the generation of molecular descriptors

The VolSurf+ software is much more user-friendly to generate molecular descriptors as almost all the prerequisite steps before descriptor calculations, including 3D structure generation, conformational analysis and geometry optimisation are performed automatically by the software, making descriptor calculation both user- and platform-independent and therefore less prone to error.

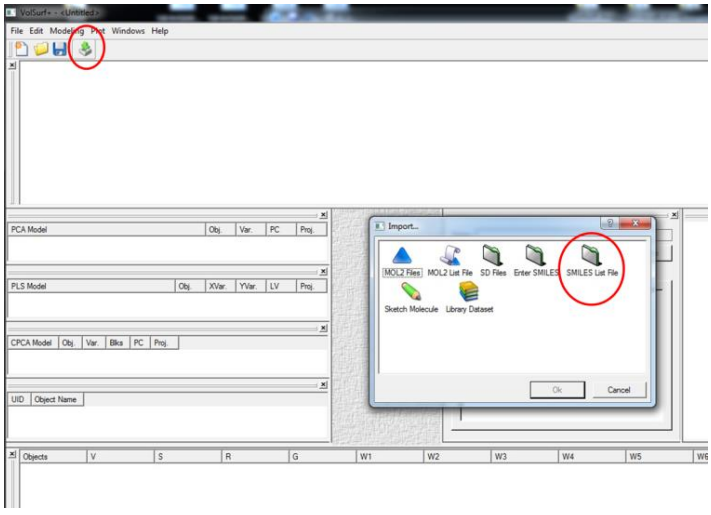
1. Extract the SMILE string for each compound in Excel and save the file in a .csv file.



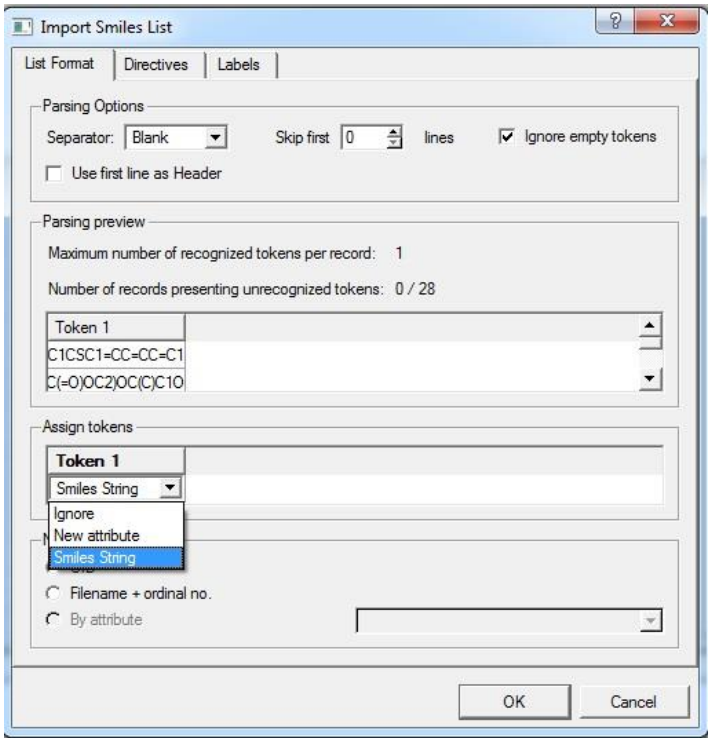
2. Open the VS+Modeller, the operation panel shows as below.



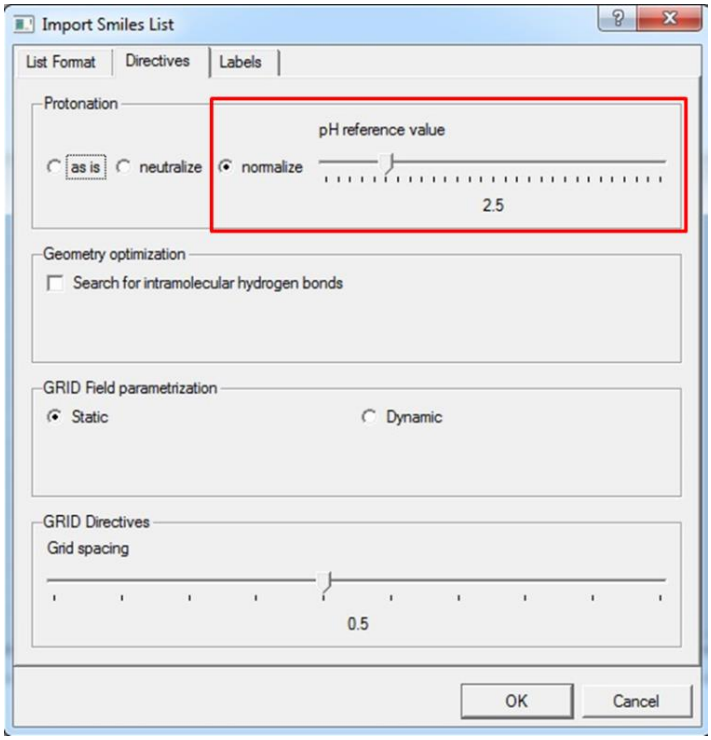
3. Click the green button to import the .csv file, then choose “SMILES List File” button.



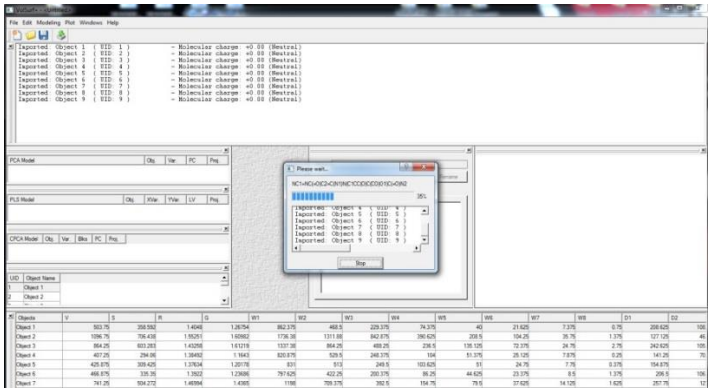
4. Import the .csv file, and click on OK icon, then select the “Smiles String” from the menu of “Assign tokens”.



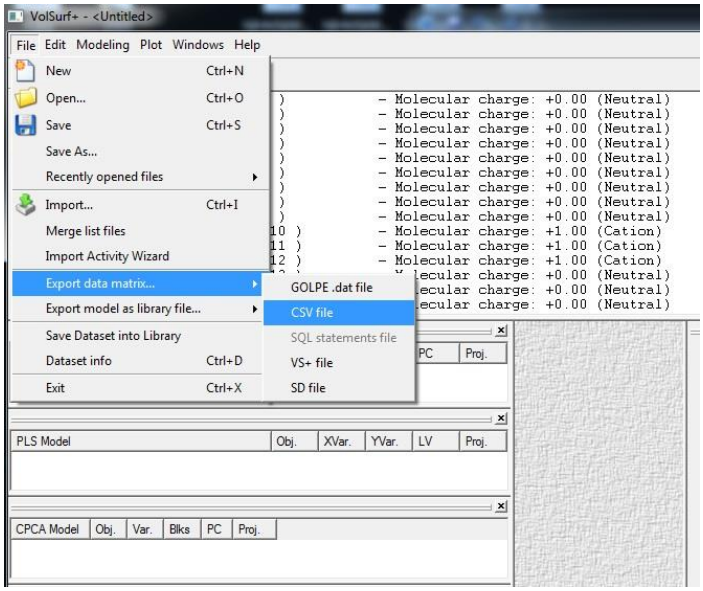
5. Go to "Directives" menu to set up the pH for the calculation of molecular descriptors, normalise a pH value as required. Then click OK icon.



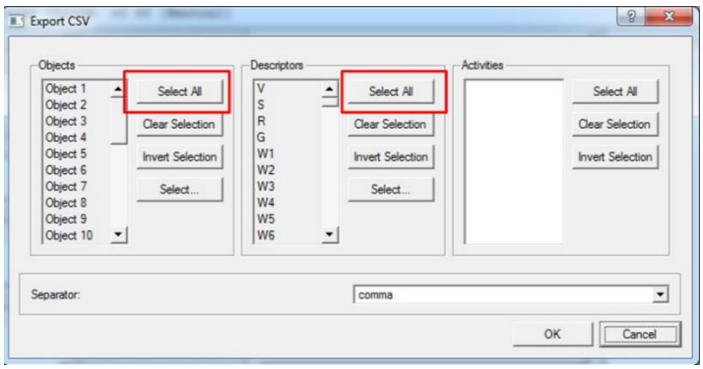
6. Then the software will start calculating the molecular descriptors, the panel shows the prepress of calculation and at the bottom you can see the generated molecular descriptors for each compound.



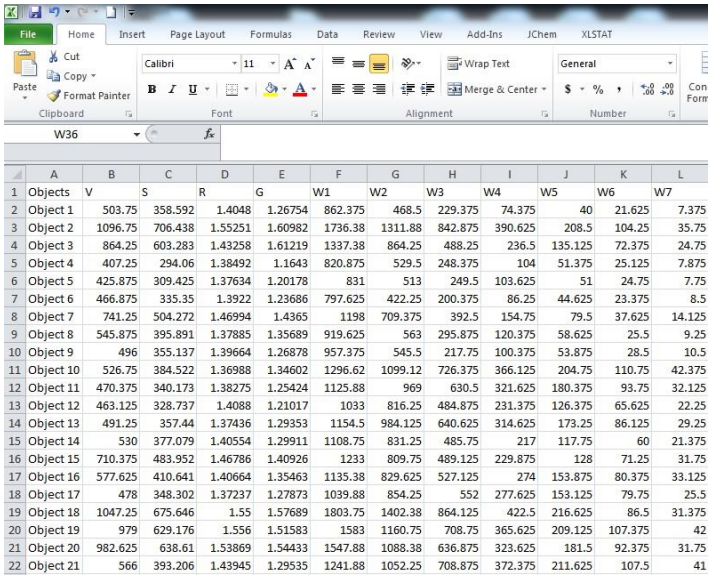
7. Once 100% finished, we need to export data matrix of molecular descriptors. Click "File" icon, then select "Export data matrix" to a file you want, here we choose "CSV file".



8. To set up the export of data matrix, select all the objects and all the descriptors, then click on OK icon to finish and save your file.



9. Lastly, check the calculated molecular descriptors, VolSurf+ generates 128 descriptors for each compound. The first column is the number of compounds, in each row, the values of molecular descriptors are listed for each compound.



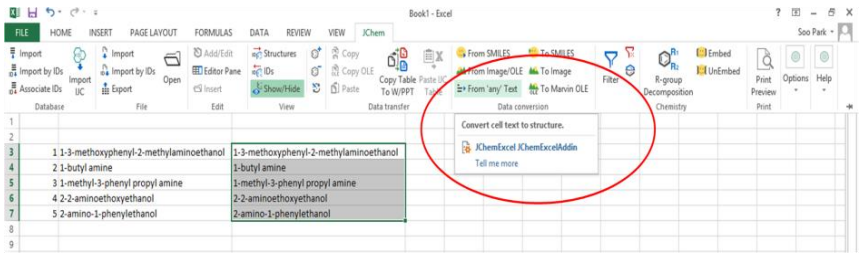
	A	B	C	D	E	F	G	H	I	J	K	L
1	Objects	V	S	R	G	W1	W2	W3	W4	W5	W6	W7
2	Object 1	503.75	358.592	1.4048	1.26754	862.375	468.5	229.375	74.375	40	21.625	7.375
3	Object 2	1096.75	706.438	1.55251	1.60982	1736.38	1311.88	842.875	390.625	208.5	104.25	35.75
4	Object 3	864.25	603.283	1.43258	1.61219	1337.38	864.25	488.25	236.5	135.125	72.375	24.75
5	Object 4	407.25	294.06	1.38492	1.1643	820.875	529.5	248.375	104	51.375	25.125	7.875
6	Object 5	425.875	309.425	1.37634	1.20178	831	513	249.5	103.625	51	24.75	7.75
7	Object 6	466.875	335.35	1.3922	1.23686	797.625	422.25	200.375	86.25	44.625	23.375	8.5
8	Object 7	741.25	504.272	1.46994	1.4365	1198	709.375	392.5	154.75	79.5	37.625	14.125
9	Object 8	545.875	395.891	1.37885	1.35689	919.625	563	295.875	120.375	58.625	25.5	9.25
10	Object 9	496	355.137	1.39664	1.26878	957.375	545.5	217.75	100.375	53.875	28.5	10.5
11	Object 10	526.75	384.522	1.36988	1.34602	1296.62	1099.12	726.375	366.125	204.75	110.75	42.375
12	Object 11	470.375	340.173	1.38275	1.25424	1125.88	969	630.5	321.625	180.375	93.75	32.125
13	Object 12	463.125	328.737	1.4088	1.21017	1033	816.25	484.875	231.375	126.375	65.625	22.25
14	Object 13	491.25	357.44	1.37436	1.29353	1154.5	984.125	640.625	314.625	173.25	86.125	29.25
15	Object 14	530	377.079	1.40554	1.29911	1108.75	831.25	485.75	217	117.75	60	21.375
16	Object 15	710.375	483.952	1.46786	1.40926	1233	809.75	489.125	229.875	128	71.25	31.75
17	Object 16	577.625	410.641	1.40664	1.35463	1135.38	829.625	527.125	274	153.875	80.375	33.125
18	Object 17	478	348.302	1.37237	1.27873	1039.88	854.25	552	277.625	153.125	79.75	25.5
19	Object 18	1047.25	675.646	1.55	1.57689	1803.75	1402.38	864.125	422.5	216.625	86.5	31.375
20	Object 19	979	629.176	1.556	1.51583	1583	1160.75	708.75	365.625	209.125	107.375	42
21	Object 20	982.625	638.61	1.53869	1.54433	1547.88	1088.38	636.875	323.625	181.5	92.375	31.75
22	Object 21	566	393.206	1.43945	1.29535	1241.88	1052.25	708.875	372.375	211.625	107.5	41

3.3 Similarity searching

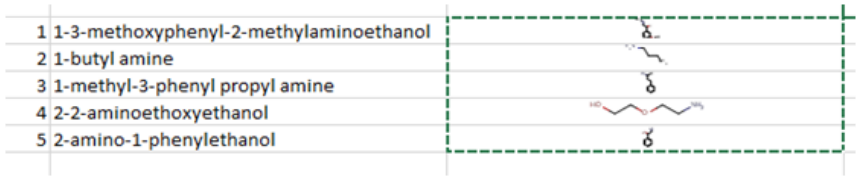
This section highlights various filtering methods used to identify appropriate training sets for QSRR modelling. In the similarity searching procedure, the target compound is matched against each compound of the database in turn, with the chosen similarity measure being used to compute the degree of resemblance in each case. The resulting set of similarity scores is then ranked in decreasing similarity order. Structural similarity searching is carried out using Tanimoto similarity coefficients (calculated using ChemAxon) with a user-defined threshold value.

Tanimoto similarity searching:

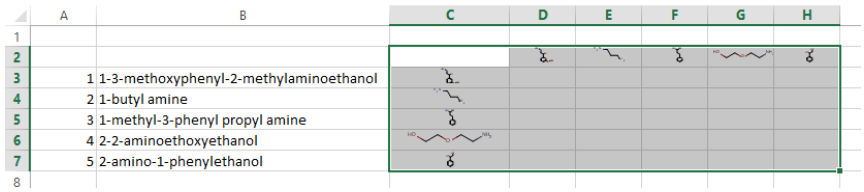
1. Prepare a Tanimoto similarity (TS) matrix for an entire dataset. In the menu "JChem", click on "From 'any' Text" to convert cell text to structure.



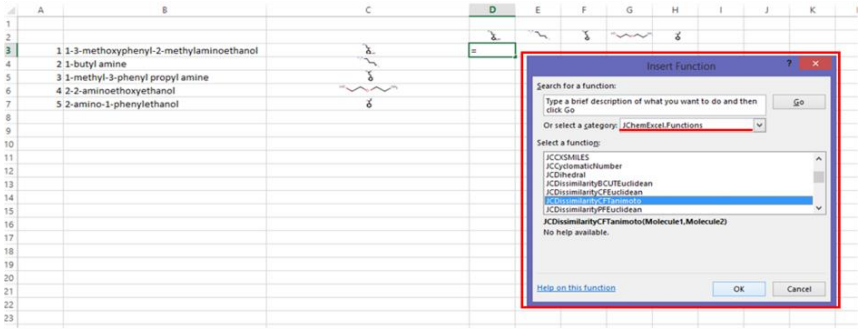
Then, the corresponding structures of compounds are converted.



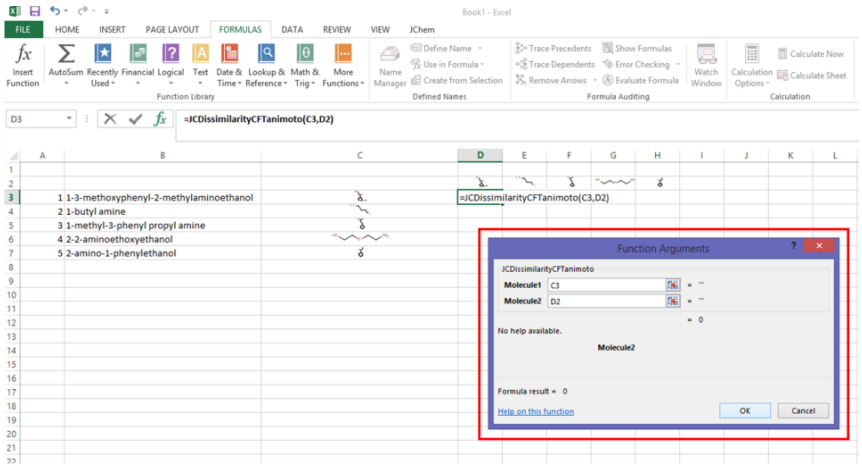
2. Prepare a matrix of compound structures for the calculation of similarity. Copy the structures of compounds from the column and then paste them in the row.



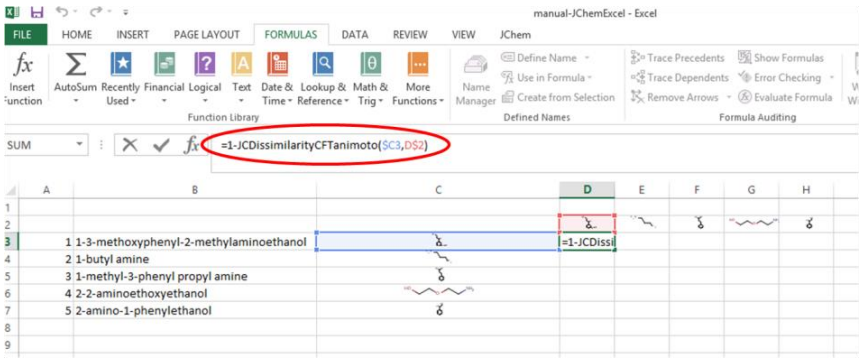
3. Open the menu “Formulas”, click on "Insert functions" and then select the category “JChemExcel”. Functions and choose the function of “JCDissimilarityCFTanimoto”.



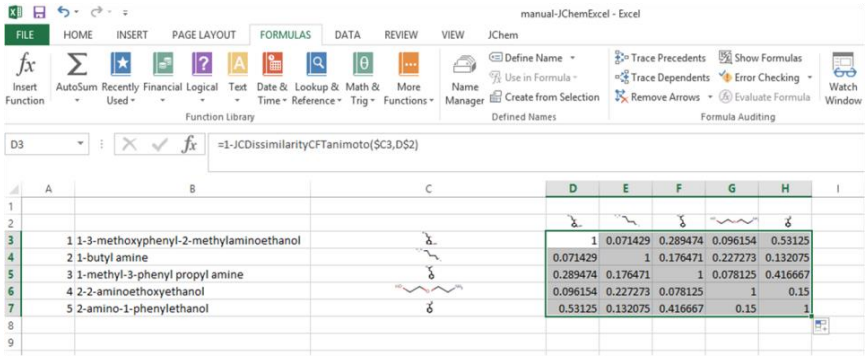
4. Define the two molecules that you want to compare (subject to pairwise Tanimoto similarity) in the pop-up windows:



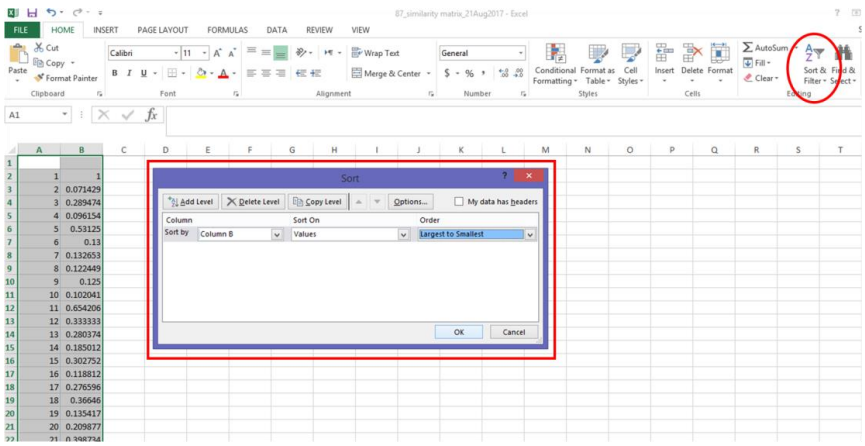
5. The function of “JCDissimilarityCFTanimoto” generates the score of dissimilarity of compounds, so here another formula "1-JCDissimilarityCFTanimoto" is used to calculate pairwise Tanimoto similarity index.



6. Using the same formula to other cells in the grid and making sure the correct compounds are compared. A similarity matrix consisting of compounds in the entire dataset is then obtained.

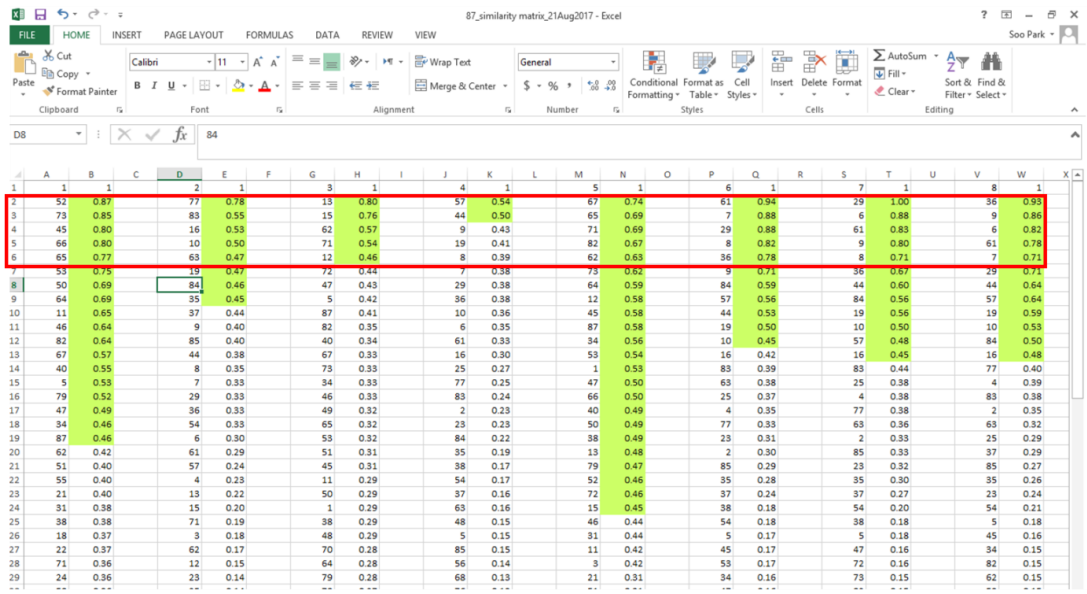


7. Each compound in the entire dataset is used sequentially as a target compound. The remaining compounds in the dataset are ranked according to their pair-wise similarity to the target. Sort the compounds based on their similarity scores as below. (the compound ID is in column A and the Tanimoto similarity values are in column B).



8. The training sets can be built using either a user-defined cut-off of similarity score, or a number of top similar compounds. For the QSRR modelling process, a minimum number of five compounds are needed for a training set. By selecting the most similar compounds or

picking compounds with TS values greater than a user-defined threshold, the training set for each target compound is prepared.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	1	1		2	1		3	1		4	1		5	1		6	1		7	1		8	1	
2	52	0.87		77	0.78		13	0.80		57	0.54		67	0.74		61	0.94		29	1.00		36	0.93	
3	73	0.85		83	0.55		15	0.76		44	0.50		65	0.69		7	0.88		6	0.88		9	0.86	
4	45	0.80		16	0.53		62	0.57		9	0.43		71	0.69		29	0.88		61	0.83		6	0.82	
5	66	0.80		10	0.50		71	0.54		19	0.41		82	0.67		8	0.82		9	0.80		61	0.78	
6	65	0.77		63	0.47		12	0.46		8	0.39		62	0.63		36	0.78		8	0.71		7	0.71	
7	53	0.75		19	0.47		72	0.44		7	0.38		73	0.62		9	0.71		36	0.67		29	0.71	
8	50	0.69		84	0.46		47	0.43		29	0.38		64	0.59		84	0.59		44	0.60		44	0.64	
9	64	0.69		35	0.45		5	0.42		36	0.38		12	0.58		57	0.56		84	0.56		57	0.64	
10	11	0.65		37	0.44		87	0.41		10	0.36		45	0.58		44	0.53		19	0.56		19	0.59	
11	46	0.64		9	0.40		82	0.35		6	0.35		87	0.58		19	0.50		10	0.50		10	0.53	
12	82	0.64		85	0.40		40	0.34		61	0.33		34	0.56		10	0.45		57	0.48		84	0.50	
13	67	0.57		44	0.38		67	0.33		16	0.30		53	0.54		16	0.42		16	0.45		16	0.48	
14	40	0.55		8	0.35		73	0.33		25	0.27		1	0.53		83	0.39		83	0.44		77	0.40	
15	5	0.53		7	0.33		34	0.33		77	0.25		47	0.50		63	0.38		25	0.38		4	0.39	
16	79	0.52		29	0.33		46	0.33		83	0.24		66	0.50		25	0.37		4	0.38		83	0.38	
17	47	0.49		36	0.33		49	0.32		2	0.23		40	0.49		4	0.35		77	0.38		2	0.35	
18	34	0.46		54	0.33		65	0.32		23	0.23		50	0.49		77	0.33		63	0.36		63	0.32	
19	87	0.46		6	0.30		53	0.32		84	0.22		38	0.49		23	0.31		2	0.33		25	0.29	
20	62	0.42		61	0.29		51	0.31		35	0.19		13	0.48		2	0.30		85	0.33		37	0.29	
21	51	0.40		57	0.24		45	0.31		38	0.17		79	0.47		85	0.29		23	0.32		85	0.27	
22	55	0.40		4	0.23		11	0.29		54	0.17		52	0.46		35	0.28		35	0.30		35	0.26	
23	21	0.40		13	0.22		50	0.29		37	0.16		72	0.46		37	0.24		37	0.27		23	0.24	
24	31	0.38		15	0.20		1	0.29		63	0.16		15	0.45		38	0.18		54	0.20		54	0.21	
25	38	0.38		71	0.19		38	0.29		48	0.15		46	0.44		54	0.18		38	0.18		5	0.18	
26	18	0.37		3	0.18		48	0.29		5	0.15		31	0.44		5	0.17		5	0.18		45	0.16	
27	22	0.37		62	0.17		70	0.28		85	0.15		11	0.42		45	0.17		47	0.16		34	0.15	
28	71	0.36		12	0.15		64	0.28		56	0.14		3	0.42		53	0.17		72	0.16		82	0.15	
29	24	0.36		23	0.14		79	0.28		68	0.13		21	0.31		34	0.16		73	0.15		62	0.15	

Retention factor-ratio (*k*-ratio) similarity searching: The ratio of retention factor for compounds in database to the retention factor of the target compound is calculated (the ratio is always ≥ 1). A *k*-ratio threshold is utilised as a similarity criteria. Compounds having a *k*-ratio smaller than the threshold are included in training set for the corresponding target compound.

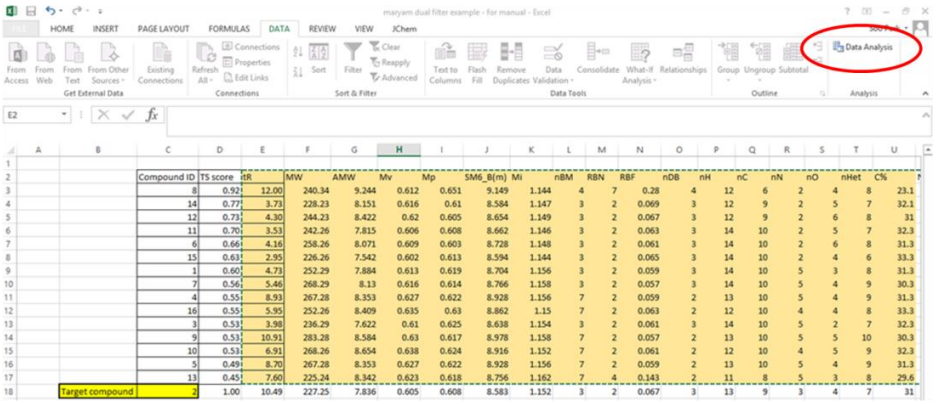
Log P/Log D similarity searching: Log P/Log D similarity searching is similar to the *k*-ratio similarity searching. The ratio of the log P (or the absolute difference of the log D) for compounds in the dataset to the log P (or log D) of the target compound is calculated. After that, compounds can be sorted based on the ratio of the log P (or the difference of the log D) values to the target. For log P or log D methods, the training set is also formed by sorting the dataset relative to the target and choosing the compounds based on the cut-off value or the top similar compounds. In the present study, a threshold of 0.2 was used in the log D approach, and a threshold of 1.2 was taken in the log P approach.

Dual filtering-based similarity searching: For the dual filtering-based similarity searching, structurally similar compounds to a target compound are clustered (the initial subset) using a TS threshold as the primary filter. Then, the most similar compound to the target is selected as a surrogate by utilising a molecular descriptor which is highly correlated to retention (coefficient of correlation > 0.8). Instead of using the target compound, the *k*-ratio similarity searching is finally applied to the surrogate compound as the secondary filter, to build a final subset (obtained from the initial subset) with chromatographically similar compounds for the QSRR modelling. The steps of dual filtering-based similarity searching are listed below.

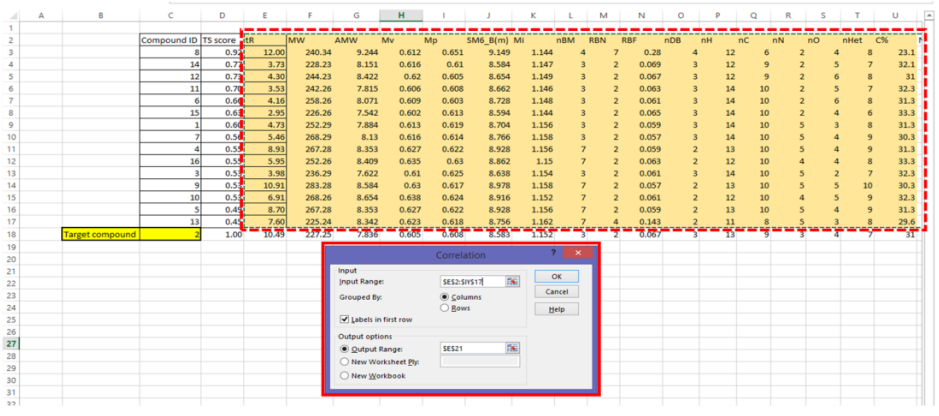
1. Load the target compound and rank compounds in the dataset based on the TS scores of compounds to the target. Find the TS subset by applying a TS threshold (using cut-off 0.45 as an example here).

	Compound ID	TS score
Target compound	2	1.00
	8	0.92
	14	0.77
	12	0.73
	11	0.70
	6	0.66
	15	0.63
	1	0.60
	7	0.56
	4	0.55
	16	0.55
	3	0.53
	9	0.53
	10	0.53
	5	0.49
	13	0.45

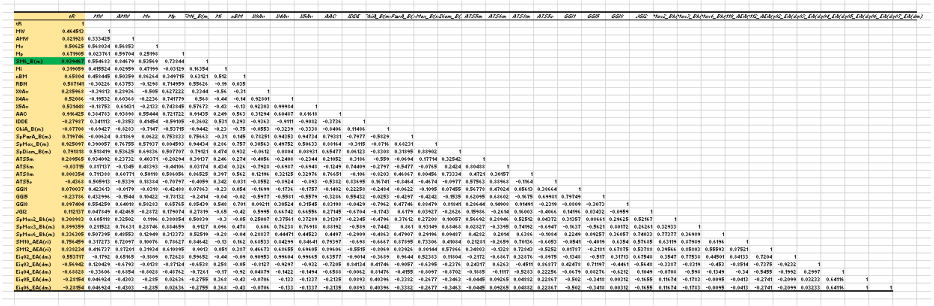
2. Finding the highly correlated (>0.8) molecular descriptor (MD) to retention time (t_R) in the initial subset using a correlation matrix. First, click on “Data Analysis”.



3. Choose “correlation” in the pop-up window and then fill in the input and output range. Check on “Labels” in first row'.



4. On the first column in the correlation matrix, find the molecular descriptor (MD) having the highest value of the correlation coefficient (r) between the MD and t_R (*e.g.*, SM6_B(m) [r=0.94]).



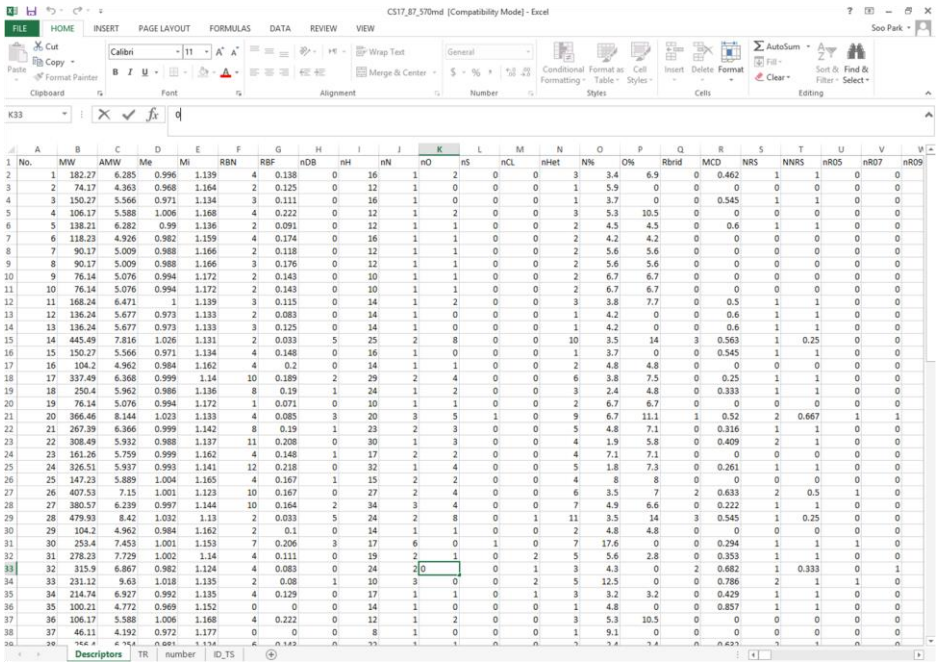
5. Calculate the absolute difference for the selected MD between the target and the compounds in the initial subset. Identify the most similar neighbour (i.e., the surrogate compound) which shows the lowest value of the absolute difference for the chosen MD. After picking the surrogate compound, apply the secondary filter: *k*-ratio similarity to the surrogate compound, to cluster compounds (*e.g.*, 14, 12, 11, 6, and 3) into the final training set. Those final training sets are used to build QSRR models and predict retentions for the target compounds.

	SM6_B(m)	0.94	<--correlation coeff.		
Compound ID	SM6_B(m)	ΔSM6_B(m)			
2	8.583		No.	k	k-ratio
8	9.149	0.566	14	0.83	1
14	8.584	0.001	8	4.87	5.90
12	8.654	0.071	15	0.44	1.87
11	8.662	0.079	12	1.10	1.34
6	8.728	0.145	11	0.73	1.13
15	8.594	0.011	6	1.03	1.25
1	8.704	0.121	16	1.91	2.32
7	8.766	0.183	1	1.32	1.59
4	8.928	0.345	7	1.67	2.02
16	8.862	0.279	4	3.37	4.08
3	8.638	0.055	3	0.95	1.15
9	8.978	0.395	5	3.26	3.95
10	8.916	0.333	9	4.34	5.25
5	8.928	0.345	10	2.38	2.89
13	8.756	0.173	13	2.72	3.29

3.4 Local QSRR modelling for single target compound

A local QSRR model provides a separate prediction of retention for each target compound, where each target compound has its own specific training set to build its own QSRR model. To perform local QSRR modelling, instructions are listed below.

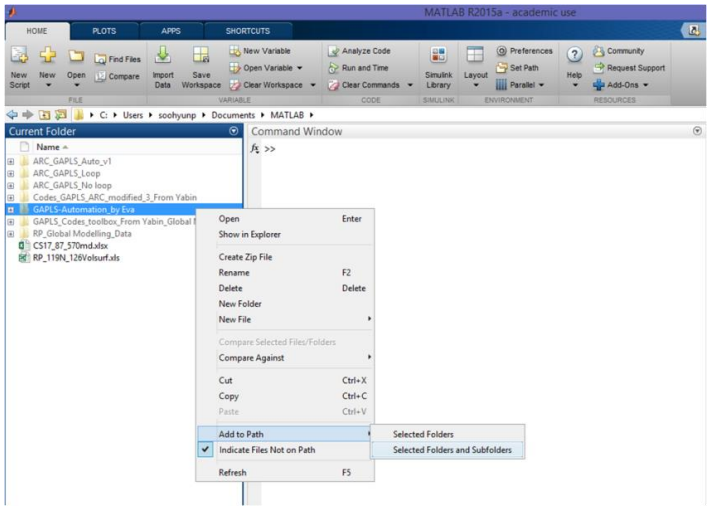
1. Prepare the dataset (*e.g.*, Excel file CS17_87_570md). The dataset consists of the following tabs (*e.g.*, Descriptors, TR, number, ID_TS) in an excel spreadsheet.



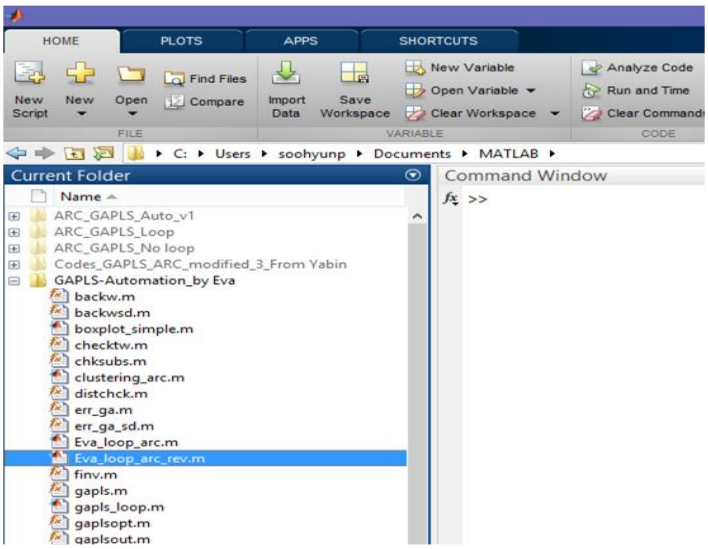
No.	MW	AMW	Me	Mi	RBN	RBF	nDB	nH	nN	nO	nS	nCl	nHet	N%	O%	Rbrid	MCD	NRS	NNRS	nR05	nR07	nR09
1	182.27	6.285	0.996	1.139	4	0.138	0	16	1	2	0	0	3	3.4	6.9	0	0.462	1	1	0	0	
2	74.17	4.363	0.968	1.164	2	0.125	0	12	1	0	0	0	1	5.9	0	0	0	0	0	0	0	
3	150.27	5.566	0.971	1.134	3	0.111	0	16	1	0	0	0	1	3.7	0	0	0.545	1	1	0	0	
4	106.17	5.588	1.006	1.168	4	0.222	0	12	1	2	0	0	3	5.3	10.5	0	0	0	0	0	0	
5	138.21	6.282	0.99	1.136	2	0.091	0	12	1	1	0	0	2	4.5	4.5	0	0.6	1	1	0	0	
6	118.23	4.926	0.982	1.159	4	0.174	0	16	1	1	0	0	2	4.2	4.2	0	0	0	0	0	0	
7	90.17	5.009	0.988	1.166	2	0.118	0	12	1	1	0	0	2	5.6	5.6	0	0	0	0	0	0	
8	90.17	5.009	0.988	1.166	3	0.176	0	12	1	1	0	0	2	5.6	5.6	0	0	0	0	0	0	
9	76.14	5.076	0.994	1.172	2	0.143	0	10	1	1	0	0	2	6.7	6.7	0	0	0	0	0	0	
10	76.14	5.076	0.994	1.172	2	0.143	0	10	1	1	0	0	2	6.7	6.7	0	0	0	0	0	0	
11	168.24	6.471	1	1.139	3	0.115	0	14	1	2	0	0	3	3.8	7.7	0	0.5	1	1	0	0	
12	136.24	5.677	0.973	1.133	2	0.083	0	14	1	0	0	0	1	4.2	0	0	0.6	1	1	0	0	
13	136.24	5.677	0.973	1.133	3	0.125	0	14	1	0	0	0	1	4.2	0	0	0.6	1	1	0	0	
14	445.49	7.816	1.026	1.131	2	0.033	5	25	2	8	0	0	10	3.5	14	3	0.563	1	0.25	0	0	
15	150.27	5.566	0.971	1.134	4	0.148	0	16	1	0	0	0	1	3.7	0	0	0.545	1	1	0	0	
16	104.2	4.962	0.984	1.162	4	0.2	0	14	1	1	0	0	2	4.8	4.8	0	0	0	0	0	0	
17	337.49	6.368	0.999	1.14	10	0.189	2	29	2	4	0	0	6	3.8	7.5	0	0.25	1	1	0	0	
18	250.4	5.962	0.986	1.136	8	0.19	1	24	1	2	0	0	3	2.4	4.8	0	0.333	1	1	0	0	
19	76.14	5.076	0.994	1.172	1	0.071	0	10	1	1	0	0	2	6.7	6.7	0	0	0	0	0	0	
20	366.46	8.144	1.023	1.133	4	0.085	3	20	3	5	1	0	9	6.7	11.1	1	0.52	2	0.667	1	1	
21	267.39	6.366	0.999	1.142	8	0.19	1	23	2	3	0	0	5	4.8	7.1	0	0.316	1	1	0	0	
22	308.49	5.932	0.988	1.137	11	0.208	0	30	1	3	0	0	4	1.9	5.8	0	0.409	2	1	0	0	
23	161.26	5.759	0.999	1.162	4	0.148	1	17	2	2	0	0	4	7.1	7.1	0	0	0	0	0	0	
24	326.51	5.937	0.993	1.141	12	0.218	0	32	1	4	0	0	5	1.8	7.3	0	0.261	1	1	0	0	
25	147.23	5.889	1.004	1.165	4	0.167	1	15	2	2	0	0	4	8	8	0	0	0	0	0	0	
26	407.53	7.15	1.001	1.123	10	0.167	0	27	2	4	0	0	6	3.5	7	2	0.633	2	0.5	1	0	
27	380.57	6.239	0.997	1.144	10	0.164	2	34	3	4	0	0	7	4.9	6.6	0	0.222	1	1	0	0	
28	479.93	8.42	1.032	1.13	2	0.033	5	24	2	8	0	1	11	3.5	14	3	0.545	1	0.25	0	0	
29	104.2	4.962	0.984	1.162	2	0.1	0	14	1	1	0	0	2	4.8	4.8	0	0	0	0	0	0	
30	253.4	7.453	1.001	1.153	7	0.206	3	17	6	0	1	0	7	17.6	0	0	0.294	1	1	1	0	
31	278.23	7.729	1.002	1.14	4	0.111	0	19	2	1	0	2	5	5.6	2.8	0	0.353	1	1	0	0	
32	315.9	6.867	0.982	1.124	4	0.083	0	24	2	0	0	1	3	4.3	0	2	0.682	1	0.333	0	1	
33	231.12	9.63	1.018	1.135	2	0.08	1	10	3	0	0	2	5	12.5	0	0	0.786	2	1	1	0	
34	214.74	6.927	0.992	1.135	4	0.129	0	17	1	1	0	1	3	3.2	3.2	0	0.429	1	1	0	0	
35	100.21	4.772	0.969	1.152	0	0	0	14	1	0	0	0	1	4.8	0	0	0.857	1	1	0	0	
36	106.17	5.588	1.006	1.168	4	0.222	0	12	1	2	0	0	3	5.3	10.5	0	0	0	0	0	0	
37	46.11	4.192	0.972	1.177	0	0	0	8	1	0	0	0	1	9.1	0	0	0	0	0	0	0	

2. Locate the dataset from the folder “Matlab”.

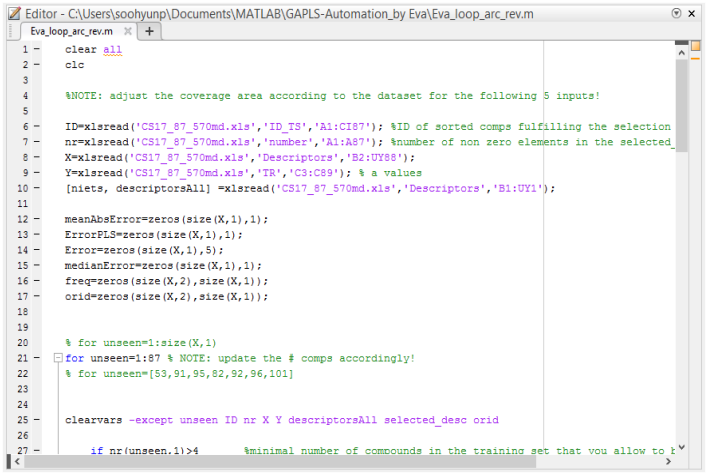
3. Launch Matlab. Right-click on the folder having the corresponding code (*i.e.*, GAPLS_Automation_by Eva) and click on “Selected Folders and Subfolders”, then click on “Add to Path”.



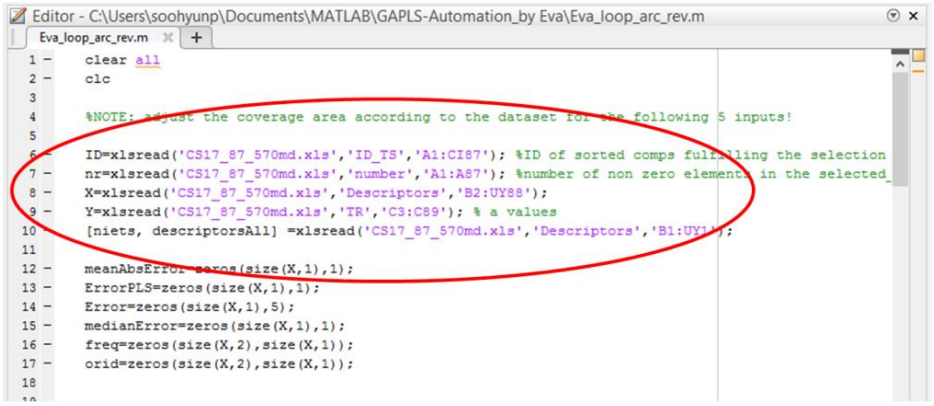
4. In the current folder, click on “Eva_loop_arc_rev.m”.



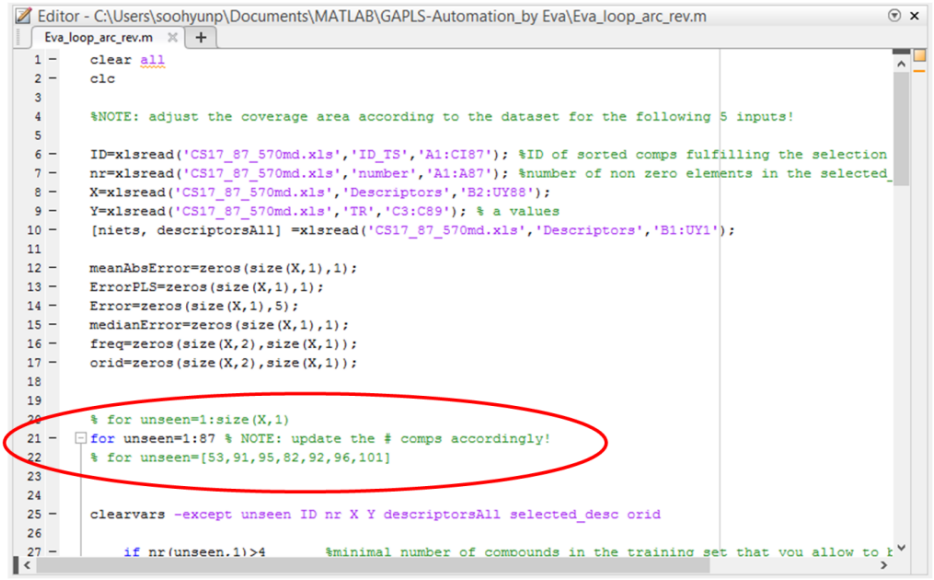
5. The editor window appears for the next input.



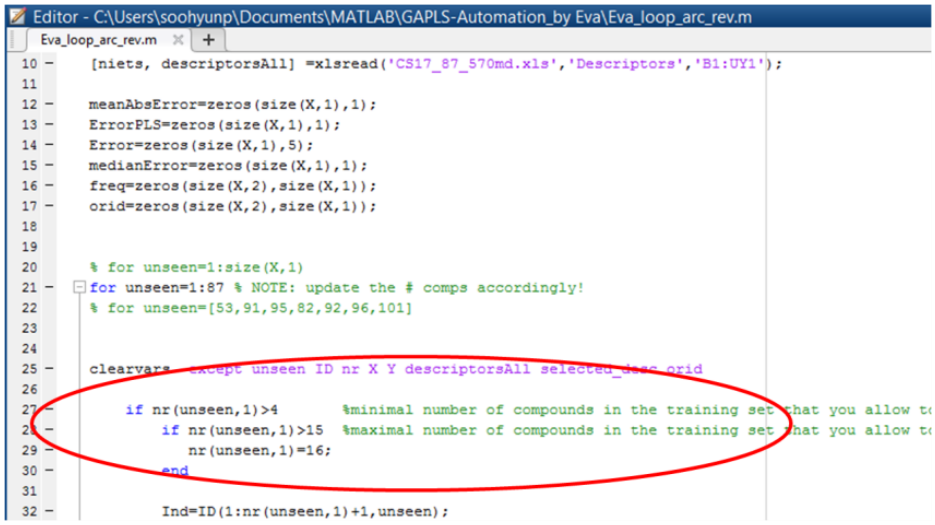
6. Adjust the coverage area according to the dataset for the following 5 inputs: ID of sorted compounds fulfilling the selection criteria (*e.g.*, TS cut-off), number of non-zero elements in the selected ID matrix, X of descriptor values, Y of response values (*e.g.* t_R values), and the names of descriptors. Input the accurate location of each piece of information in that Excel file to make sure the modelling process can be performed smoothly.



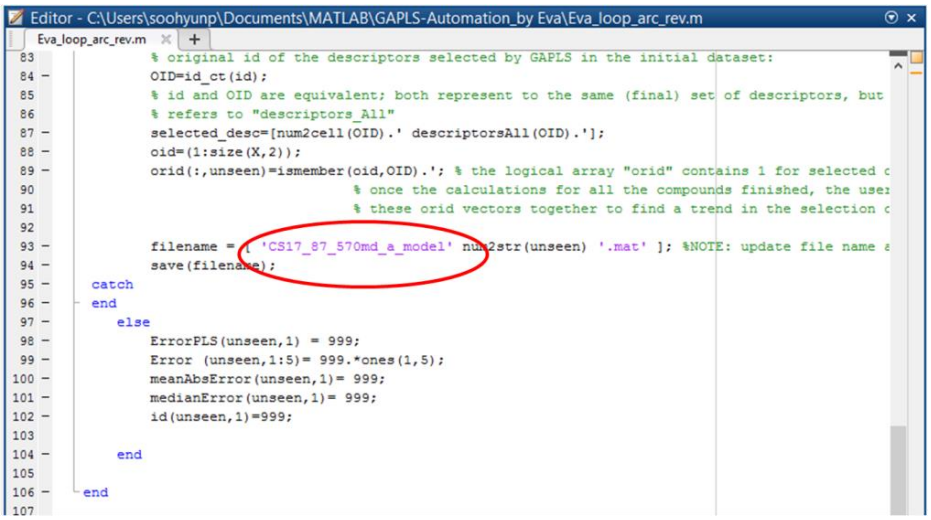
7. Type the ID numbers of compounds (*e.g.* 1:87) to be modelled.



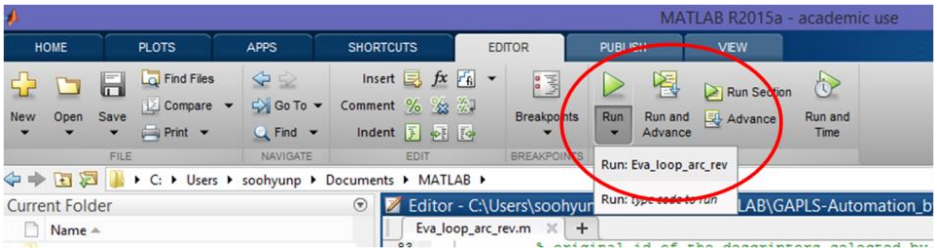
8. Define the minimal and maximal number of compounds in the training set to build the model, the minimum number of compounds in a training set is five.



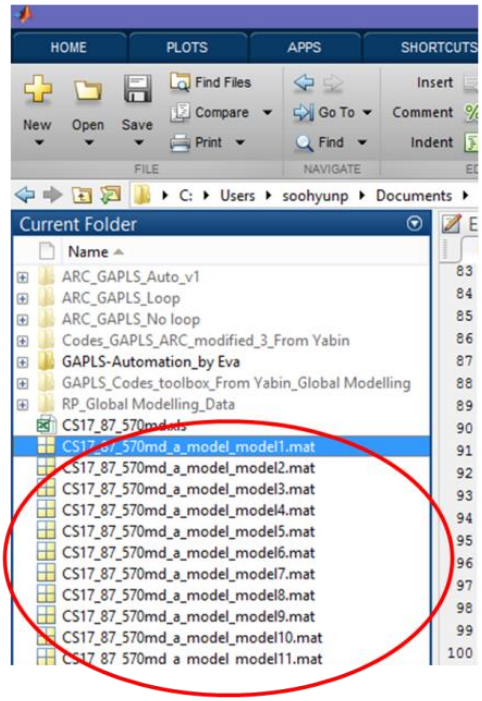
9. Update the file name accordingly.



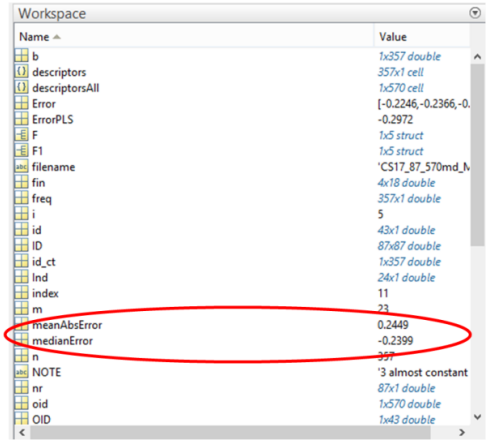
10. Then, run the code: "Eva_loop_arc_rev".



11. Once finished, the generated models will appear in the current folder in Matlab. Click on the file (*e.g.*, CS17_87_570md_a_model1.mat) in the current folder, which shows the result for the model of compound ID1.



12. The result for the model 1 will appear in the workspace. Statistics like the mean absolute error (MAE), median Error and others are shown. For further error calculations, choose an appropriate error reporting according to the conditions.

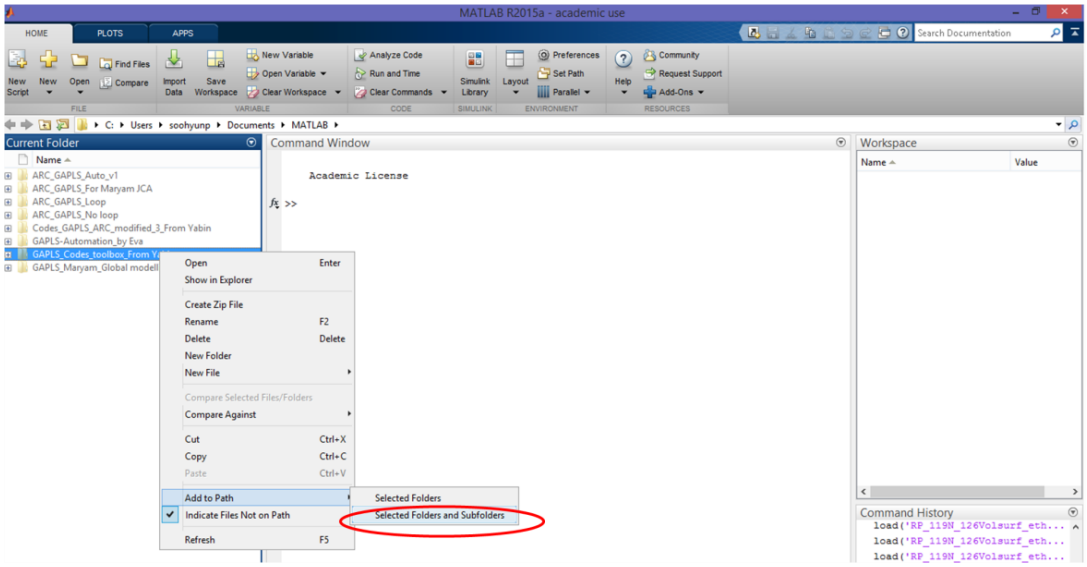


3.5 Local QSRR modelling for a group of target compounds

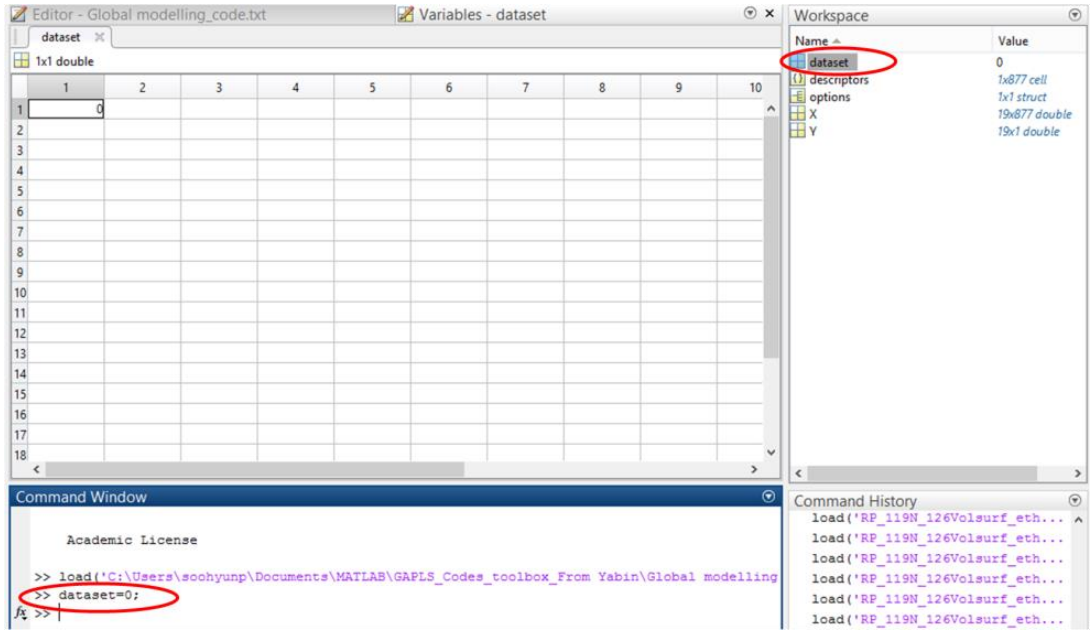
A global QSRR model is one which can predict the retention times for many target compounds (more than one compound) at once. On the contrary, local QSRR model can be built ether for an individual target compound, or a group of test compounds. Two types of filtering approaches were employed during the QSRR modelling to yield training set for the prediction of a group of test compounds.

Local Compound Type (LCT) filtering: compounds are grouped into three clusters based on their chemical nature: acids (that can donate a proton), bases (that can accept a proton), and neutral compounds. Here, 199 neutral compounds are taken as an example.

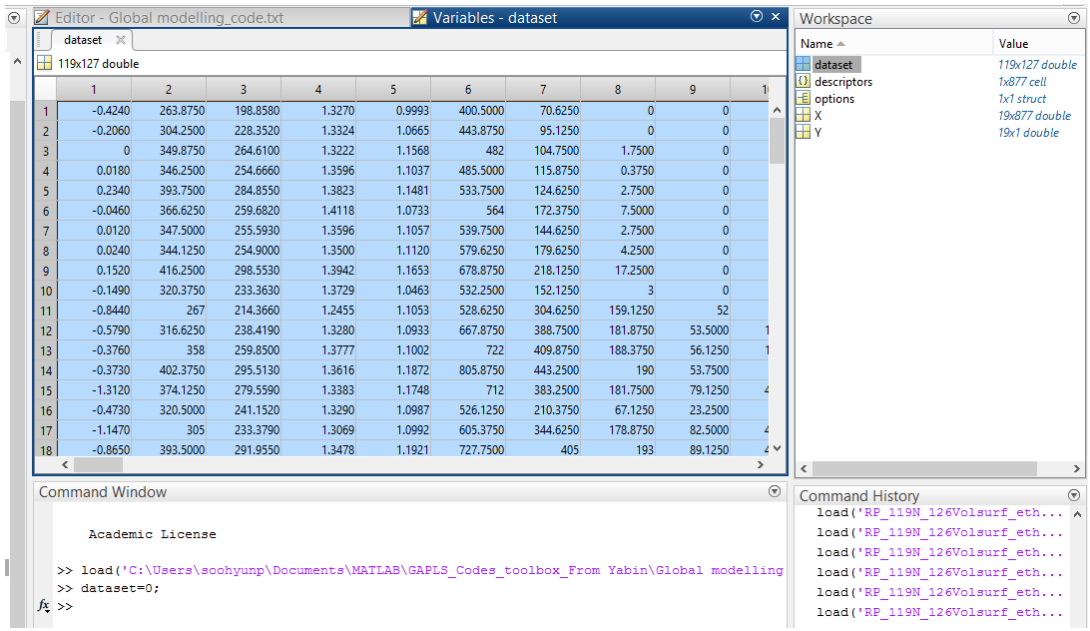
1. Launch Matlab. Right-click on the folder having the corresponding code (*i.e.*, GAPLS_Codes_toolbox_From Yabin) and click on “Selected Folders and Subfolders” on the “Add to Path”.



2. Click on "modelling.mat" in the current folder, which contains necessary parameters (dataset, descriptors, options, X and Y) for the modelling in the workspace (see the figure below). First, click on "dataset" in workspace and type "dataset=0;" in "Command Window".

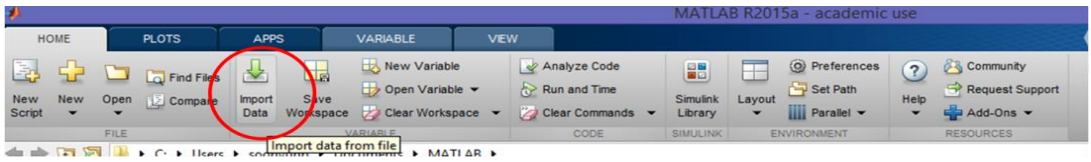


3. Input the dataset (consisting of retention parameters for the neutral compounds and their molecular descriptors) from the Excel spreadsheet and then paste it on the worksheet in the editor window.

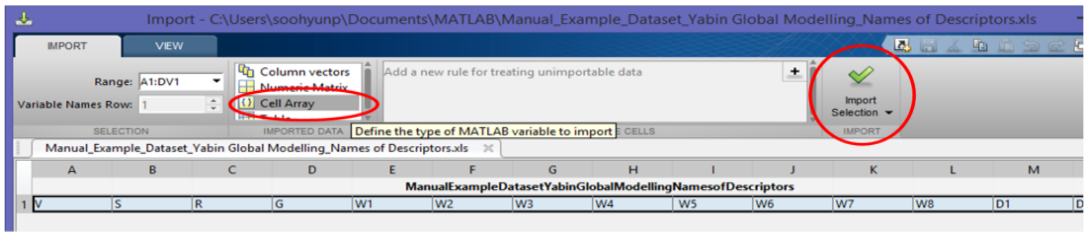


4. Input the data for X matrix and Y matrix following the same procedure. Here, X represents independent variables in the modelling process (should be molecular descriptors of compounds), Y represents dependent parameters in the modelling process (should be retention times of compounds).

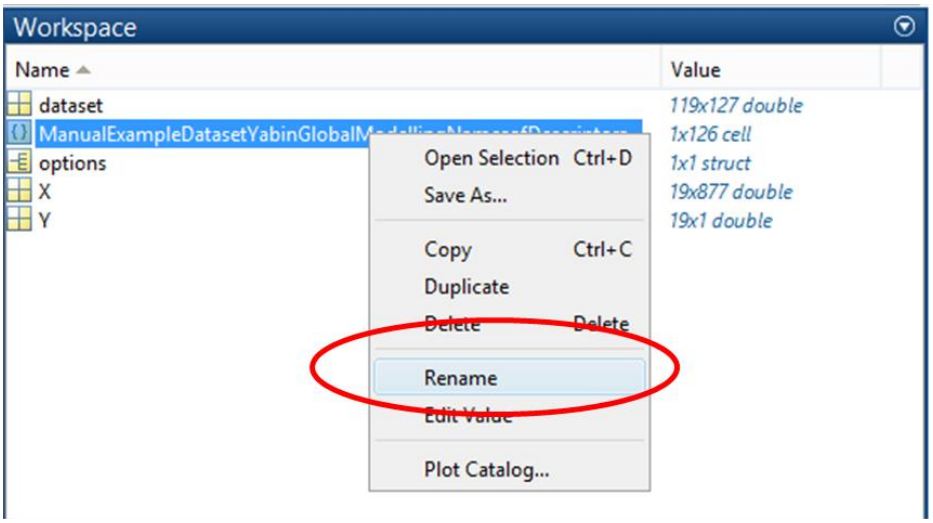
5. Input the names of molecular descriptors, click on "Import Data" to upload the Excel file which contains the names of molecular descriptors in a row.



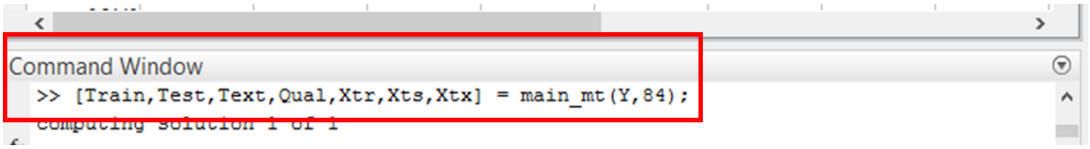
6. The names of molecular descriptors are different from numbers, therefore upload the names as a "Cell Array", and then click on "Import Selection".



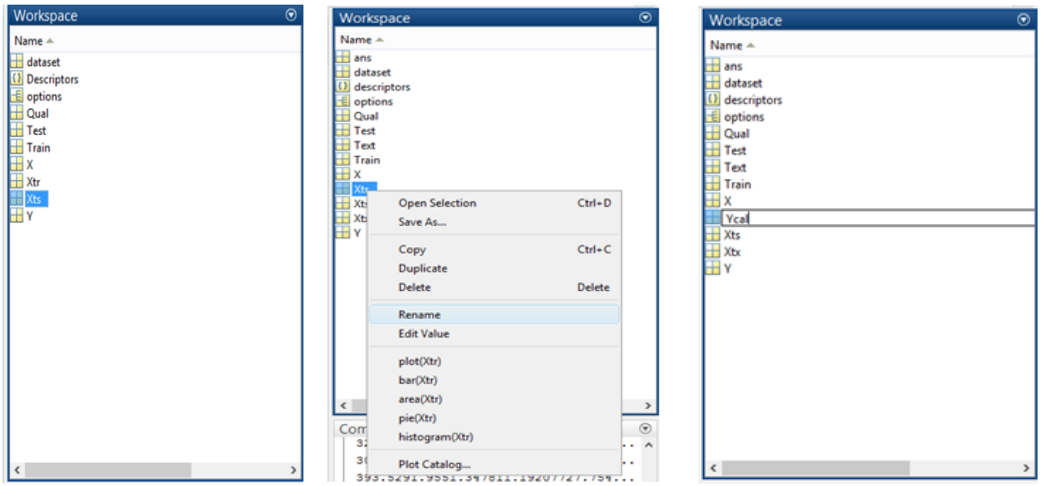
7. Rename the uploaded cell array.



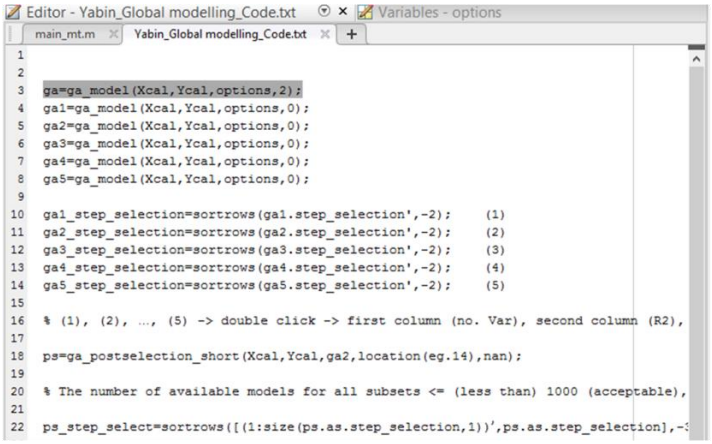
8. To generate training and test sets using the D-optimal algorithm, use the code "[Train, Test, Text, Qual, Xtr, Xts, Xtx] = main_mt (Y, dim);" in the command window. 70% of compounds in each cluster were selected as the training set and the remaining compounds (30%) were used as a test set. Y is the matrix which contains of retention times for 119 neutral compounds, 84 is the number of compounds that will be allocated into the training set.



9. After running the D-optimal code, the ID of compounds in training set and test set are given in the "Train" and "Test" matrix, respectively. Meanwhile, the responses (retention times) of compounds in training and test set are also generated automatically and saved in the "Xtr" and "Xts" matrix. Rename these two matrixes as "Ycal" and "Ytest" in the workspace for subsequent use.



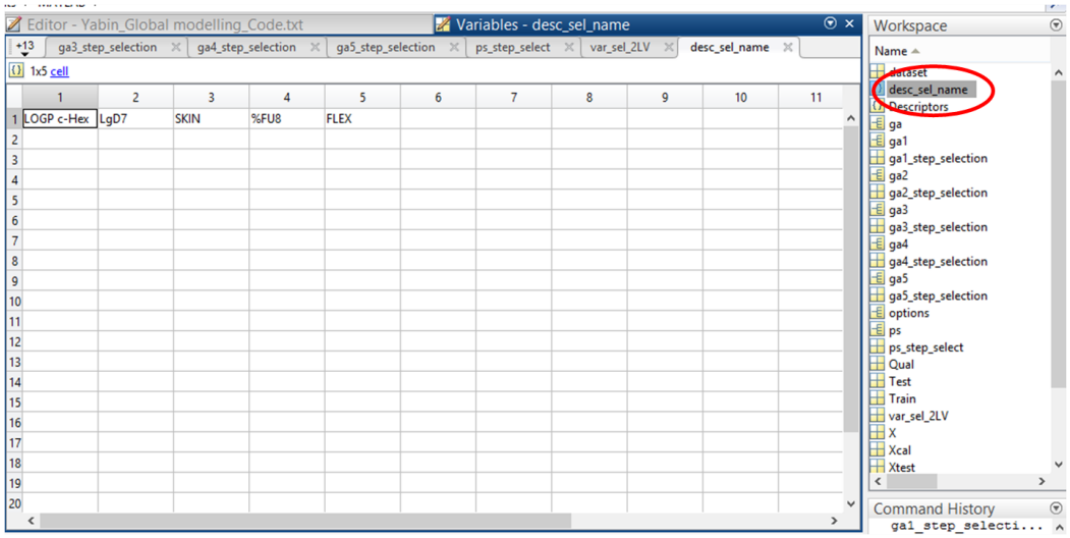
10. Having the ID of compounds for training and test set, now the corresponding variables (molecular descriptors) are ready to generate by running the code of "Xcal = X (Train,:);" and "Xtest=X(Test,:);" in command window. The Xcal matrix contains the values of molecular descriptors for training compounds, where the Xtest matrix include the values of molecular descriptors for test compounds. Then, the GA-PLS modelling can be performed. Click on the file "modelling_Code.txt" in the current folder and then the instruction file appears in the editor window.



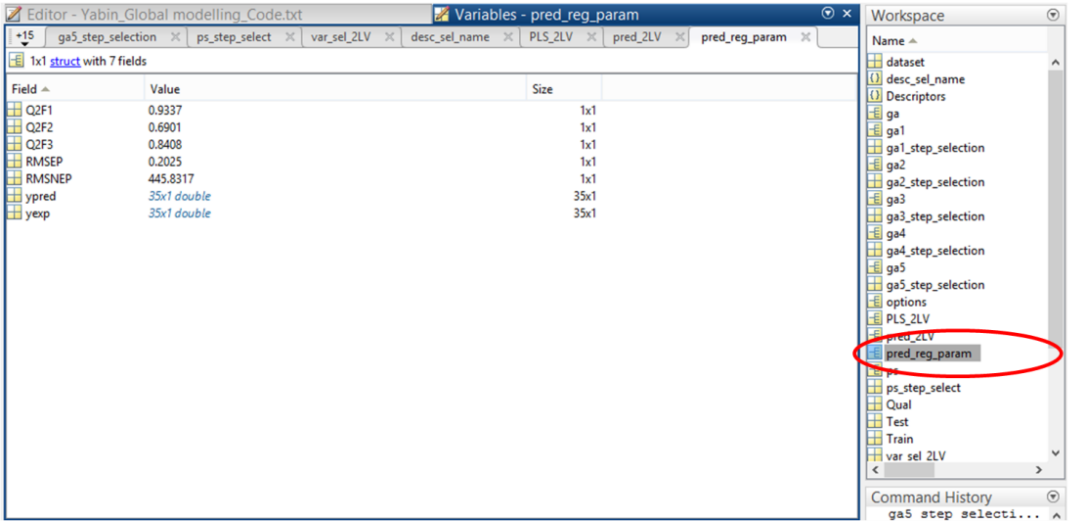
11. Follow the instructions for running the GA process, (*e.g.*, ga=ga_model(Xcal, Ycal, options,2);) by pasting the code in the command window.

12. The GA process is run five times to choose the most informative descriptors for the construction of QSRR models. After that, PLS is used to build QSRR models by running the code of "pred_nLV = plstest(Xtest(:,var_sel_nlv),PLS_nLV);". n is the number of latent variables (LV), here an optimum number of LV need to be entered.

13. The PLS model gives the prediction of retention for compounds in both training and test sets, as well as the selected descriptors used to build the QSRR models, to review the selected descriptors, click on the "desc_sel_name" matrix generated in the workspace.



14. To see the performance of the QSRR model, click on "pred_reg_param" matrix in the workspace to review the statistic parameters including the RMSE, and the predicted retention for training and test compounds.



Local Second Dominant Interaction (LSDI) filtering: using this method compounds are clustered based on their secondary dominant interaction after hydrophobicity, namely:

Hydrophobic interaction only – η'

Steric bulk – σ'

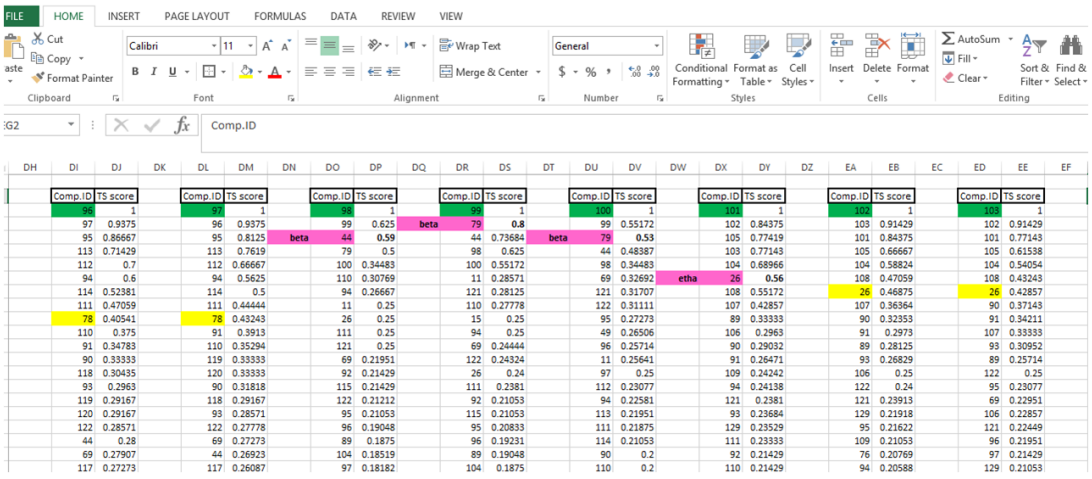
Hydrogen bonding basicity– β'

Hydrogen bonding acidity – α'

Ionic interaction (charge) – κ'

More than one type of secondary interaction - cluster 6

Each of these six clusters was modelled individually by utilising approximately 70% of its compounds (using the D-optimal approach as above) as the training set and the rest as the external test set. To allocate new test compounds into the corresponding cluster so that the right model can be utilised, Tanimoto similarity (TS) searching was introduced. The structural similarity of each new compound was investigated against the training compounds in each cluster with the aim of finding one training compound with a highest pairwise Tanimoto structural similarity score (at least 0.5). For example, a compound with a highest Tanimoto score against training compounds is around 0.41 (less than 0.5), so this compound will be excluded in the modelling. If such a compound is found, this target compound will be assigned into the same LDSI cluster, allowing the correct model is used for its retention prediction. Compounds will be excluded if their pairwise TS indices are less than 0.5 when calculated against compounds in the training set.



Comp.ID	TS score	Comp.ID	TS score	Comp.ID	TS score	Comp.ID	TS score	Comp.ID	TS score	Comp.ID	TS score	Comp.ID	TS score	Comp.ID	TS score
96	1	97	1	98	1	99	1	100	1	101	1	102	1	103	1
97	0.9375	96	0.9375	99	0.625	beta	79	0.8	99	0.55172	102	0.84375	103	0.91429	102	0.91429
95	0.86667	95	0.8125	beta	44	0.59	44	0.73684	beta	79	0.53	105	0.77419	101	0.84375	101	0.77143
113	0.71429	113	0.7619	79	0.5	98	0.625	44	0.48387	103	0.77143	105	0.66667	105	0.61538
112	0.7	112	0.66667	100	0.34483	100	0.55172	98	0.34483	104	0.68966	104	0.58824	104	0.54054
94	0.6	94	0.5625	110	0.30769	11	0.28571	69	0.32692	etha	26	0.56	108	0.47059	108	0.43243
114	0.52381	114	0.5	94	0.26667	121	0.28125	121	0.31707	108	0.55172	26	0.46875	26	0.42857
111	0.47059	111	0.44444	11	0.25	110	0.27778	122	0.31111	107	0.42857	107	0.36364	90	0.37143
78	0.40541	78	0.43243	26	0.25	15	0.25	95	0.27273	89	0.33333	90	0.32353	91	0.34211
110	0.375	91	0.3913	111	0.25	94	0.25	49	0.26506	106	0.2963	91	0.2973	107	0.33333
91	0.34783	110	0.35294	121	0.25	69	0.24444	96	0.25714	90	0.29032	89	0.28125	93	0.30952
90	0.33333	119	0.33333	69	0.21951	122	0.24324	11	0.25641	91	0.26471	93	0.26829	89	0.25714
118	0.30435	120	0.33333	92	0.21429	26	0.24	97	0.25	109	0.24242	106	0.25	122	0.25
93	0.2963	90	0.31818	115	0.21429	111	0.2381	112	0.23077	94	0.24138	122	0.24	95	0.23077
119	0.29167	118	0.29167	122	0.21212	92	0.21053	94	0.22581	121	0.2381	121	0.23913	69	0.22951
120	0.29167	93	0.28571	95	0.21053	115	0.21053	113	0.21951	93	0.23684	129	0.21918	106	0.22857
122	0.28571	122	0.27778	96	0.19048	95	0.20833	111	0.21875	129	0.23529	95	0.21622	121	0.22449
44	0.28	69	0.27273	89	0.1875	96	0.19231	114	0.21053	111	0.23333	109	0.21053	96	0.21951
69	0.27907	44	0.26923	104	0.18519	89	0.19048	90	0.2	92	0.21429	76	0.20769	97	0.21429
117	0.27273	117	0.26087	97	0.18182	104	0.1875	110	0.2	110	0.21429	94	0.20588	129	0.21053

The training set was used to build QSRR models in the same way as outlined above using the GA-PLS method in Matlab, the predictive ability of the models was verified using test sets.

3.6 Sum of ranking difference analysis

The overall performance of the generated models using different filtering approaches can be compared using the sum of ranking difference (SRD) approach. In this work, objects including R^2, Q^2, the %RMSEP of the test set, and the slope of the regression line were selected and evaluated. Moreover, the number of test compounds, the average number of descriptors used for modelling, the number of constructed models for each approach and the prediction frequency with an absolute error less than 30s and 60s were also considered. A given benchmark was used as the gold standard. The program can be run using a "SRDrep_V5_E10" script on the Excel file. The instructions for performing a SRD analysis are listed below.

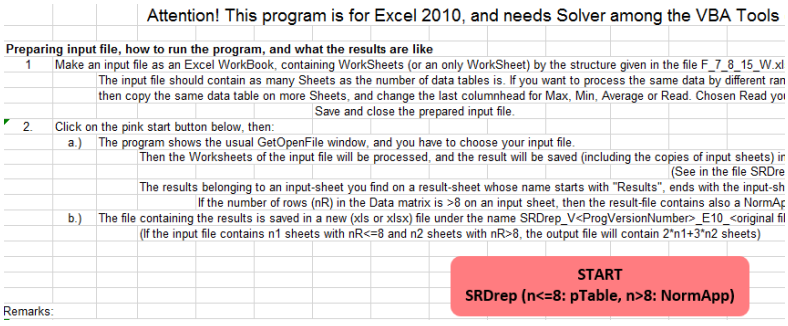
1. Prepare the input matrix with objects (parameters) and variables (models) for the SRD analysis.

Analyte	G126	G34	LTS	LLD	LCT	LSDI	Gold Standard
R^2	0.7005	0.5395	0.9189	0.6760	0.8240	0.9232	0.9943
Q^2	0.6844	0.5523	0.9100	0.6697	0.8173	0.9234	0.9922
slope	0.6284	0.6096	1.0097	0.6695	0.7629	0.9450	1.0000
RMSEP%	24.27	31.56	17.53	36.87	23.18	20.79	4.37
no. of test compounds	44	44	51	119	49	55	148
no. of descriptors used	126	34	45	22	7	7	7
no. of models constructed	1	1	51	119	3	6	1
Prediction frequency with absolute error less than 30 s	0.49	0.49	0.49	0.67	0.71	0.71	0.94
Prediction frequency with absolute error less than 60 s	0.68	0.71	0.79	0.64	0.79	0.83	0.97

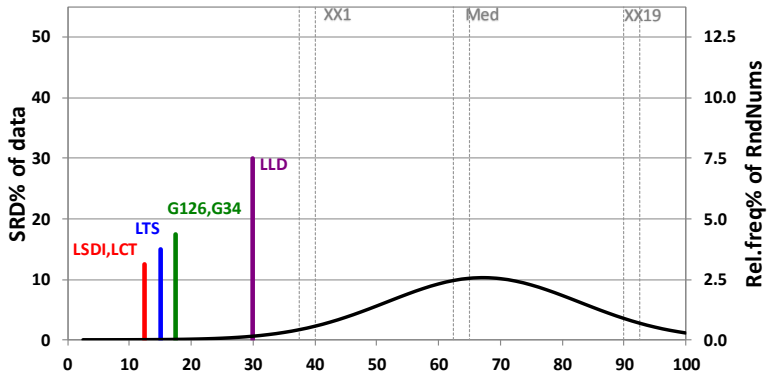
2. Make an input file as an Excel file, containing worksheets by a given structure as below. Put the values of gold standard in the "Read" column.

	A	B	C	D	E	F	G	H
1	n_row=	9			n_col=	6		
2	WD_Ex7				n_dig=	2		
3		G126	G34	LTS	LLD	LCT	LSDI	Read
4	F1	0.6293	0.4573	0.9189	0.6066	0.8017	0.9219	0.9943
5	F2	0.6844	0.5523	0.9100	0.6697	0.8173	0.9234	0.9922
6	F3	0.7921	0.8056	1.0037	0.8398	0.8653	0.9712	1.0000
7	F4	24.2700	31.5600	17.5300	36.8700	23.1800	20.7900	4.3700
8	F5	44.0000	44.0000	51.0000	119.0000	49.0000	55.0000	148.0000
9	F6	126.00	34.00	45.00	22.00	7.00	7.00	7.00
10	F7	0.49	0.49	0.49	0.67	0.71	0.71	0.94
11	F8	1	1	51	119	3	6	1
12	F9	0.68	0.71	0.8	0.64	0.79	0.83	0.98

3. Click on the pink start button below, then the program shows the usual GetOpenFile window, click on it to input the above-mentioned file. Once uploaded, click on the next to run the program.



4. Results can be reviewed in another worksheet in the same Excel file. From the SRD graph and the ranking table, the validity of the generated models and the best model are evaluated.



References

1. Hall, L.M., D.W. Hill, L.C. Menikarachchi, M.-H. Chen, L.H. Hall, and D.F. Grant, *Optimizing artificial neural network models for metabolomics and systems biology: an example using HPLC retention index data.* Bioanalysis, 2015. **7**(8): p. 939-955.
2. Tyteca, E., M. Talebi, R. Amos, S.H. Park, M. Taraji, Y. Wen, R. Szucs, C.A. Pohl, J.W. Dolan, and P.R. Haddad, *Towards a chromatographic similarity index to establish localized quantitative structure-retention models for retention prediction: use of retention factor ratio.* Journal of Chromatography A, 2017. **1486**: p. 50-58.
3. Wen, Y., M. Talebi, R.I. Amos, R. Szucs, J.W. Dolan, C.A. Pohl, and P.R. Haddad, *Retention prediction in reversed phase high performance liquid chromatography using quantitative structure-retention relationships applied to the Hydrophobic Subtraction Model.* Journal of Chromatography A, 2018. **1541**: p. 1-11.